Quantitative Economics 12 (2021), 77–108 1759-7331/20210077

Sensitivity analysis using approximate moment condition models

Timothy B. Armstrong
Department of Economics, Yale University

Michal Kolesár
Department of Economics, Princeton University

We consider inference in models defined by approximate moment conditions. We show that near-optimal confidence intervals (CIs) can be formed by taking a generalized method of moments (GMM) estimator, and adding and subtracting the standard error times a critical value that takes into account the potential bias from misspecification of the moment conditions. In order to optimize performance under potential misspecification, the weighting matrix for this GMM estimator takes into account this potential bias and, therefore, differs from the one that is optimal under correct specification. To formally show the near-optimality of these CIs, we develop asymptotic efficiency bounds for inference in the locally misspecified GMM setting. These bounds may be of independent interest, due to their implications for the possibility of using moment selection procedures when conducting inference in moment condition models. We apply our methods in an empirical application to automobile demand, and show that adjusting the weighting matrix can shrink the CIs by a factor of 3 or more.

Keywords. Sensitivity analysis, confidence intervals, misspecification, generalized method of moments, semiparametric efficiency.

JEL classification. C12, C13, C52.

1. Introduction

Economic models are typically viewed as approximations of reality. However, conventional approaches to estimation and inference assume that a model holds exactly. In this paper, we weaken this assumption, and consider inference in a class of models characterized by moment conditions which are only required to hold in an approximate sense. The failure of the moment conditions to hold exactly may come from failure of exclusion restrictions (e.g., through omitted variable bias or because instruments enter the

Timothy B. Armstrong: timothy.armstrong@yale.edu
Michal Kolesár: mkolesar@princeton.edu

We thank Isaiah Andrews, Gary Chamberlain, Tim Christensen, Mikkel Plagborg-Møller, Jesse Shapiro, Martin Weidner and participants at several conferences and seminars for helpful comments and suggestions, and Soonwoo Kwon for research assistance. All remaining errors are our own. Armstrong acknowledges support by National Science Foundation Grant SES-1628939. Kolesár acknowledges support by the Sloan Research Fellowship, and by the National Science Foundation Grant SES-1628878.

 https://doi.org/10.3982/QE1609

structural equation directly in an IV model), functional form misspecification, or other sources such as measurement error, or data contamination.

We assume that we have a model characterized by a set of population moment conditions $g(\theta)$. In the generalized method of moments (GMM) framework, for instance, $g(\theta) = E[g(w_i, \theta)]$, which can be estimated by the sample analog $\frac{1}{n}\sum_{i=1}^n g(w_i, \theta)$, based on the sample $\{w_i\}_{i=1}^n$. When evaluated at the true parameter value θ_0, the population moment condition lies in a known set specified by the researcher,

$$g(\theta_0) = c/\sqrt{n}, \quad c \in \mathcal{C}.$$

The set $\mathcal{C}$ formalizes the way in which the moment conditions may fail, and it can then be varied as a form of sensitivity analysis, with $\mathcal{C} = \{0\}$ reducing to the correctly specified case.

We focus on local misspecification: the scaling by $\sqrt{n}$ implies that the specification error and the sampling error are of the same order of magnitude, and it ensures that the asymptotic approximation captures the fact that it may not be clear from the sample at hand whether the model is correctly specified. It also leads to increased tractability, allowing us to deliver a simple method for inference on a structural parameter of interest $h(\theta_0)$, rather than a pseudo-true parameter. This tractability has made local misspecification a popular tool for sensitivity analysis in applied work, especially following the recent influential paper by Andrews, Gentzkow, and Shapiro (2017).[1] As with any asymptotic device, our modeling of misspecification as local should not be taken to mean that we literally believe that the model would be closer to correct if we had more data. Rather, its usefulness should be judged by whether it yields accurate approximations to the finite-sample behavior of estimators and confidence intervals, which in our case requires that the set $\mathcal{C}/\sqrt{n}$ be small relative to sampling uncertainty.

We propose a simple method for constructing asymptotically valid confidence intervals (CIs) under this setup: one takes a standard estimator, such as the GMM estimator, and adds and subtracts its standard error times a critical value that takes into account the potential asymptotic bias of the estimator, in addition to its variance. A key insight of this paper is that because the CIs must be widened to take into account the potential bias, the optimal weighting matrix for the correctly specified case (the inverse of the variance matrix of the moments) is generally no longer optimal under local misspecification. Rather, the optimal weighting matrix takes into account potential misspecification in the moments in addition to the variance of their estimates: it places less weight on moments that are allowed to be further from zero according the researcher's specification of the set $\mathcal{C}$. We also show that an analogous result holds for other performance criteria, such as estimation under the mean-squared error: the optimal weighting matrix again trades off the potential misspecification of the moments against their precision, although the optimal tradeoff is different.

To illustrate the practical importance of this result, we apply our methods to form misspecification-robust CIs in an empirical model of automobile demand based on

[1] For recent empirical examples using local sensitivity analysis, see Gayle and Shephard (2019) or Duflo, Greenstone, Pande, and Ryan (2018).

Berry, Levinsohn, and Pakes (1995). We consider sets $\mathcal{C}$ motivated by the forms of local misspecification considered in Andrews, Gentzkow, and Shapiro (2017), who calculate the asymptotic bias of the usual GMM estimator in this model. We find that adjusting the weighting matrix to account for potential misspecification substantially reduces the potential bias of the estimator and, as a result, leads to large efficiency improvements of the optimal CI relative to a CI based on the GMM estimator that is optimal under correct specification: it shrinks the CI by up to a factor of 3 or more in our main specifications. As a result, we obtain informative CIs in this model even under moderate amounts of misspecification.

We show that the CIs we propose are near-optimal when the set $\mathcal{C}$ is convex and centrosymmetric ($c \in \mathcal{C}$ implies $-c \in \mathcal{C}$). To this end, we argue that the relevant "limiting experiment" for the locally misspecified GMM model is isomorphic to an approximately linear model of Sacks and Ylvisaker (1978), which falls under a general framework studied by, among others, Donoho (1994), Cai and Low (2004) and Armstrong and Kolesár (2018). We derive asymptotic efficiency bounds for CIs in the locally misspecified GMM model that formally translate bounds from the approximately linear limiting experiment to the locally misspecified GMM setting. In particular, these bounds imply that the scope for improvement over our CIs by optimizing expected length at a particular value of θ_0 and $c = 0$ (while still maintaining coverage over the whole parameter space for θ and $\mathcal{C}$) is limited, even if one optimizes expected length at the true values of θ_0 and c.

These efficiency bounds address an important criticism of our CIs: they require a priori specification of the set $\mathcal{C}$ that defines misspecification, including both the *magnitude* of misspecification and *which* moments are misspecified. In particular, one cannot substantively improve upon our CI by, say, trying to use data-driven methods that gauge misspecification magnitude or try to determine which moments are misspecified. These bounds have implications for procedures proposed by Andrews and Guggenberger (2009), DiTraglia (2016) and McCloskey (2020), who consider the case where some moments are known to be correctly specified and no a priori bound is placed on the magnitude of misspecification of the remaining moments. As we discuss in Section 4.3.2, in this case our CI reduces to the usual CI based on the k_1 correctly specified moments, and our efficiency bounds show that CIs proposed in these papers cannot substantively improve upon it.

Because we cannot use the data to determine the magnitude M of the set $\mathcal{C}$, we recommend plotting our CIs as a function of the potential misspecification magnitude M, or reporting the smallest value of M for which a particular finding breaks down. Such sensitivity analysis is easy to conduct under out proposed implementation. In particular, we show that, when the set $\mathcal{C}$ is characterized by ℓ_p constraints, the class of weightings that trace out the optimal bias-variance tradeoff as a function of how much relative weight we put on the bias can be easily computed by recasting the problem as a penalized regression problem. By exploiting this analogy, we develop a simple algorithm for computing this class under ℓ_∞ constraints that is similar to the LASSO/LAR algorithm (Efron, Hastie, Johnstone, and Tibshirani (2004), Rosset and Zhu (2007)); under ℓ_2 constraints, the solution admits a closed form.[2] Furthermore, as we discuss in Sec-

[2]An R package implementing our CIs under ℓ_p constraints is available at https://github.com/kolesarm/GMMSensitivity.

tion 3, this class of weightings is entirely determined by the shape of $\mathcal{C}$; its magnitude M only determines the optimal relative weight we should put on the bias. Thus, tracing out the optimal weighting as a function of M can be done at essentially no additional computational cost. Furthermore, to avoid having to reoptimize the objective function with respect to the new weighting matrix, one can also form the CIs by adding and subtracting our critical value from a one-step estimator (see Newey and McFadden (1994, Section 3.4)) based on any initial estimate that is $\sqrt{n}$-consistent under correct specification. We illustrate this approach in our empirical application in Section 6.

Our paper is related to several strands of literature. Our efficiency results are related to those in Chamberlain (1987) for point estimation in the correctly specified setting (see also Hansen (1985)) and, more broadly, semiparametric efficiency theory in correctly specified settings (see, e.g., Chapter 25 in van der Vaart (1998)). As we discuss in Section 4.3, some of our efficiency results are novel even in the correctly specified case, and may be of independent interest. Kitamura, Otsu, and Evdokimov (2013) considered efficiency of point estimators satisfying certain regularity conditions when the misspecification is bounded by the Hellinger distance. As we discuss in more detail in Section 4.3.4, our results imply that under this form of misspecification, the optimal weighting matrix remains the same as under correct specification; both the usual GMM estimator and the estimator proposed by Kitamura, Otsu, and Evdokimov (2013) can thus be used to form near-optimal CIs, and both estimators have the same local asymptotic minimax properties.

Local misspecification has been used in a number of papers, which include, among others, Newey (1985), Berkowitz, Caner, and Fang (2012), Conley, Hansen, and Rossi (2012), Guggenberger (2012), and Bugni and Ura (2019), and has antecedents in the literature on robust statistics (see Huber and Ronchetti (2009), and references therein). Andrews, Gentzkow, and Shapiro (2017) considered this setting and note that asymptotic bias of a regular estimator can be calculated using influence function weights, which they call the sensitivity, and show how such calculations can be used for sensitivity analysis in applications (see also extensions of these ideas in Andrews, Gentzkow, and Shapiro 2020 and Mukhin 2018). Our results imply that, if one is interested in inference, conclusions of such sensitivity analysis may be substantially sharpened by using the misspecification-optimal weighting matrix, or, equivalently, the misspecification-optimal sensitivity. In independent work, Bonhomme and Weidner (2020) provide a framework for estimation and inference in misspecified likelihood models when the misspecification set $\mathcal{C}$ is defined with respect to a larger class of models using statistical notions of distance. Our focus is on overidentified moment condition models, as in Andrews, Gentzkow, and Shapiro (2017), and we are agnostic about how $\mathcal{C}$ is determined. The proposal to use estimators that optimize an asymptotic bias-variance tradeoff using the influence function is common to both papers. The efficiency bounds in Section 4 are unique to the present paper.

The rest of this paper is organized as follows. Section 2 presents our misspecification robust CIs. Section 3 gives step-by-step instructions for computing our CIs, along with discussion of other practical issues. Section 4 presents efficiency bounds for CIs in

locally misspecified models; it can be skipped by readers interested only in implementing the methods. Section 5 discusses applications to particular moment condition models. Section 6 presents an empirical application. Additional results and proofs are collected in the Appendices in the Online Supplementary Material (Armstrong and Kolesár (2021)).

2. Misspecification-robust CIs

We have a model that maps a vector of parameters $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$ to a d_g-dimensional population moment condition $g(\theta)$ that restricts the distribution of the observed data $\{w_i\}_{i=1}^n$. We allow the moment condition model to be locally misspecified, so that at the true value θ_0, the population moment condition is not necessarily zero, but instead lies in a $\sqrt{n}$-neighborhood of 0:

$$g(\theta_0) = c/\sqrt{n}, \quad c \in \mathcal{C}, \tag{1}$$

where $\mathcal{C} \subseteq \mathbb{R}^{d_g}$ is a known set. The set $\mathcal{C}$ may allow for misspecification in potentially all moment conditions; we do not require that some elements of c are zero. Our goal is to construct a CI for a scalar $h(\theta_0)$, where $h\colon \mathbb{R}^{d_\theta} \to \mathbb{R}$ is a known function. For example, if we are interested in one of the elements θ_j of θ, we would take $h(\theta) = \theta_j$. More generally, the function h will be nonlinear, as is, for example, generally the case when θ is a vector of supply or demand parameters, and $h(\theta)$ is an elasticity, or some counterfactual.

This setup allows (but does not require) both θ_0 and $h(\theta_0)$ to have the same interpretation as in the correctly specified case, so that our CIs may still be interpreted as CIs for the structural parameter, elasticity, or counterfactual of interest. For this interpretation, one typically needs to rule out forms of misspecification that affect the mapping $\theta \mapsto h(\theta)$. While we do not formally consider cases in which this mapping itself is misspecified, such cases are covered under a mild generalization of our framework, in which h is a function of both θ and c.

Note that the interpretation of $h(\theta_0)$, and the conceptual framework defining θ_0 is not affected by our modeling of the misspecification as local: given a set $\widetilde{\mathcal{C}}_n = \mathcal{C}/\sqrt{n}$, the moment conditions $g(\theta_0) \in \widetilde{\mathcal{C}}_n$ describe the restrictions that the data generating process and the researcher's modeling assumptions place on θ_0.[3] The plausibility of these restrictions is evaluated for a given sample size at hand; it does not depend on assumptions about how $\widetilde{\mathcal{C}}_n$ changes with n. While we focus on sequences $\widetilde{\mathcal{C}}_n = \mathcal{C}/\sqrt{n}$, we discuss in Remark 3.3 how our insights can be used to construct CIs that are valid global misspecification, when $\widetilde{\mathcal{C}}_n$ is fixed with n.

To formalize the notion of asymptotic validity and efficiency of CIs, we will need to allow the true parameter value θ_0 as well as the vector c and the data generating process (and hence the map $\theta \mapsto g(\theta)$) to vary with the sample size. For clarity of exposition, we

[3]Formally, for a given sample size n, θ_0 may be set identified, and the identified set under a distribution P is defined as the set of parameters θ_0 that satisfy the moment conditions $E_P g(w_i, \theta_0) \in \widetilde{\mathcal{C}}_n$ where E_P denotes expectation under P. We construct CIs that cover $h(\theta_0)$ for points θ_0 in the identified set (see Imbens and Manski (2004), for a discussion of this notion of coverage). See Section 4 and Appendix C for formal definitions of coverage and optimality of our CIs.

focus here on the case in which these parameters are fixed. See Theorem 4.1 and Appendix C for the general case. Under some forms of misspecification, such as functional form misspecification, there may be additional higher-order terms on the right-hand side of (1); our results remain unchanged if this is the case. Again, for clarity of exposition, we focus on the case in which (1) holds exactly.

We assume that the sample moment condition $\hat{g}(\theta)$, constructed using the data $\{w_i\}_{i=1}^n$, satisfies

$$\sqrt{n}\big(\hat{g}(\theta_0) - g(\theta_0)\big) \xrightarrow{d} \mathcal{N}(0, \Sigma), \tag{2}$$

where $\xrightarrow{d}$ denotes convergence in distribution as $n \to \infty$. In the GMM model, the population and sample moment conditions are given by $g(\theta) = E[g(w_i, \theta)]$ and $\hat{g}(\theta) = \frac{1}{n}\sum_{i=1}^n g(w_i, \theta)$, respectively, where $g(\cdot, \cdot)$ is a known function. However, to cover other minimum distance problems, we do not require that the moment conditions necessarily take this form. We further assume that the moment condition is smooth enough so that

$$\text{for any } \theta_n = \theta_0 + \mathcal{O}_P(1/\sqrt{n}), \quad \hat{g}(\theta_n) - \hat{g}(\theta_0) = \Gamma(\theta_n - \theta_0) + o_P(1/\sqrt{n}), \tag{3}$$

where Γ is the $d_g \times d_\theta$ derivative matrix of g at θ_0. Conditions (2) and (3) are standard regularity conditions in the literature on linear and nonlinear estimating equations; see Newey and McFadden (1994) for primitive conditions. Finally, we also assume that h is continuously differentiable with the $1 \times d_\theta$ derivative matrix at θ_0 given by H.

2.1 *CIs based on asymptotically linear estimators*

Under correct specification, when $\mathcal{C} = \{0\}$, standard estimators $\hat{h}$ of $h(\theta)$ are asymptotically linear in $\hat{g}(\theta_0)$. This will typically extend to our locally misspecified case, so that for some vector $k \in \mathbb{R}^{d_g}$,

$$\sqrt{n}\big(\hat{h} - h(\theta_0)\big) = k'\sqrt{n}\hat{g}(\theta_0) + o_P(1) \xrightarrow{d} \mathcal{N}\big(k'c, k'\Sigma k\big), \tag{4}$$

where the convergence in distribution follows by (1) and (2). If in addition, the estimator is regular (so that equality in equation (4) holds uniformly for θ in a $\sqrt{n}$-neighborhood of θ_0), then k will satisfy (see, e.g., Section 2 in Newey (1990))

$$H = -k'\Gamma. \tag{5}$$

For example, in a GMM model, if we take $\hat{h} = h(\hat{\theta}_W)$ where

$$\hat{\theta}_W = \operatorname*{argmin}_\theta \hat{g}(\theta)' W \hat{g}(\theta), \tag{6}$$

is the GMM estimator with weighting matrix W, equations (4) and (5) will hold with $k' = -H(\Gamma' W \Gamma)^{-1}\Gamma' W$ (see Newey (1985)). Because the vector k determines the local asymptotic bias of the estimator, we follow Andrews, Gentzkow, and Shapiro (2017), and refer to k as the *sensitivity* of $\hat{h}$.

We now show how to construct misspecification-robust CIs based on an asymptotically linear estimator $\hat{h}$ with a given sensitivity k. In Section 2.2, we show how to choose this sensitivity optimally, to achieve the shortest CI among those based on reg-

ular asymptotically linear estimators. In Section 4, we will show that, under this choice of k, the resulting CI is (near) optimal not only within the class of CIs based on regular asymptotically linear estimators, but among *all* CIs that satisfy the asymptotic coverage requirement.

Let $\hat{k}$ and $\hat{\Sigma}$ be consistent estimates of k and Σ. Then by Slutsky's theorem,

$$\frac{\sqrt{n}\big(\hat{h} - h(\theta_0)\big)}{\sqrt{\hat{k}'\hat{\Sigma}\hat{k}}} \xrightarrow{d} \mathcal{N}\Big(\frac{k'c}{\sqrt{k'\Sigma k}}, 1\Big).$$

Under correct specification, the right-hand side corresponds to a standard normal distribution, and we can form a CI with asymptotic coverage $100 \cdot (1-\alpha)\%$ as $\hat{h} \pm z_{1-\alpha/2}\sqrt{\hat{k}'\hat{\Sigma}\hat{k}/n}$, where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of a $\mathcal{N}(0,1)$ distribution; this is the usual Wald CI.

When we allow for misspecification, the Wald CI will no longer be valid. However, note that the asymptotic bias $k'c/\sqrt{k'\Sigma k}$ is bounded in absolute value by $\overline{\text{bias}}_{\mathcal{C}}(k)/\sqrt{k'\Sigma k}$ where $\overline{\text{bias}}_{\mathcal{C}}(k) \equiv \sup_{c\in\mathcal{C}}|k'c|$. Therefore, given c, the z-statistic in the preceding display is asymptotically $\mathcal{N}(t,1)$ where $|t| \leq \overline{\text{bias}}_{\mathcal{C}}(k)/\sqrt{k'\Sigma k}$. This leads to the CI

$$\hat{h} \pm \text{cv}_\alpha\Big(\frac{\overline{\text{bias}}_{\mathcal{C}}(\hat{k})}{\sqrt{\hat{k}'\hat{\Sigma}\hat{k}}}\Big) \cdot \sqrt{\hat{k}'\hat{\Sigma}\hat{k}}/\sqrt{n}, \tag{7}$$

where $\text{cv}_\alpha(\bar{t})$ is the $1-\alpha$ quantile of $|Z|$, with $Z \sim \mathcal{N}(\bar{t},1)$. In particular, $\text{cv}_\alpha(0) = z_{1-\alpha/2}$, so that in the correctly specified case, (7) reduces to the usual Wald CI. As we discuss in Section 4, in the limiting experiment, this CI becomes equivalent to the fixed-length CI proposed by Donoho (1994).

To form a one-sided CI based on an estimator $\hat{h}$ with sensitivity k, one can simply subtract its maximum bias, in addition to the standard error:

$$[\hat{h} - \overline{\text{bias}}_{\mathcal{C}}(\hat{k})/\sqrt{n} - z_{1-\alpha}\sqrt{\hat{k}'\hat{\Sigma}\hat{k}/n}, \infty). \tag{8}$$

One could also form a valid two-sided CI by adding and subtracting the worst-case bias $\overline{\text{bias}}_{\mathcal{C}}(\hat{k})/\sqrt{n}$ from $\hat{h}$, in addition to adding and subtracting $z_{1-\alpha/2}\sqrt{\hat{k}'\hat{\Sigma}\hat{k}/n}$; however, since $\hat{h}$ cannot simultaneously have a large positive and a large negative bias, such CI will be conservative, and longer than the CI in (7).

2.2 *Optimal CIs*

The asymptotic length of the CI in equation (7) is given by

$$2 \cdot \text{cv}_\alpha\big(\overline{\text{bias}}_{\mathcal{C}}(k)/\sqrt{k'\Sigma k}\big) \cdot \sqrt{k'\Sigma k}/\sqrt{n}. \tag{9}$$

To attain the shortest possible CI, we therefore need to use an estimator with sensitivity that minimizes this expression. We restrict attention to asymptotically linear estimators that are regular, so that we need to minimize (9) subject to (5). The CI length in equa-

tion (9) depends on θ only through Σ. Furthermore, it depends on the sensitivity only through the maximum bias, $\overline{\text{bias}}_{\mathcal{C}}(k)$, and the variance $k'\Sigma k$. Therefore, rather than minimizing (9) directly over all sensitivities k, one can first minimize the variance subject to a bound $\overline{B}$ on the worst-case bias,

$$\min_k k'\Sigma k \quad \text{s.t. (5)} \quad \text{and} \quad \sup_{c\in\mathcal{C}} |k'c| \le \overline{B}, \tag{10}$$

and then vary the bound $\overline{B}$ to find the bias-variance trade-off that leads to the shortest CI. In our implementation in Section 3, we focus on the case where $\mathcal{C}$ is characterized by ℓ_p constraints, in which case a closed-form expression for the worst-case bias $\sup_{c\in\mathcal{C}}|k'c|$ is available, and it is computationally trivial to solve (10) directly or in Lagrange multiplier form. In general, when the set $\mathcal{C}$ is convex, one can reformulate (10) as a convex optimization problem, leading to a computationally tractable solution (see Section 4). One can also use (10) to determine the optimal sensitivity for constructing one-sided CIs, if we use quantiles of excess length as the criterion for choosing a CI. We provide details in Appendix C.

Once the optimal sensitivity has been determined, we can implement an estimator with this sensitivity as a one-step estimator. In particular, let $\hat{\theta}_{\text{initial}}$ be an initial $\sqrt{n}$-consistent estimator of θ_0, let $\hat{k} = k + o_P(1)$ be a consistent estimator of the desired sensitivity k. Then the one-step estimator

$$\hat{h} = h(\hat{\theta}_{\text{initial}}) + \hat{k}'\hat{g}(\hat{\theta}_{\text{initial}})$$

will have the desired sensitivity. This follows from the Taylor expansion

$$\begin{aligned}\sqrt{n}\big(\hat{h} - h(\theta_0)\big) &= H\sqrt{n}(\hat{\theta}_{\text{initial}} - \theta_0) + \hat{k}'\sqrt{n}\hat{g}(\hat{\theta}_{\text{initial}}) + o_P(1)\\ &= \big(H + \hat{k}'\Gamma\big)\sqrt{n}(\hat{\theta}_{\text{initial}} - \theta_0) + \hat{k}'\sqrt{n}\hat{g}(\theta_0) + o_P(1),\end{aligned}$$

where the second line follows from (3). It then follows from (5) that the first term converges in probability to zero, and $\hat{h}$ satisfies (4).

3. Practical implementation

We now give step-by-step instructions for computing our CI. To make it easy to determine the sensitivity of the CI to the magnitude of misspecification, we consider sets of the form $\mathcal{C} = \mathcal{C}(M) = \{Mc\colon c \in \mathcal{C}(1)\}$, where the scalar M measures the magnitude of misspecification. We discuss the exact specification of the set $\mathcal{C}(M)$ in Remark 3.1 below.

The fact that M simply scales the potential magnitude of misspecification leads to a simplification when tracing out the optimal CI as a function of M. In particular, let $\{k_\lambda\}_{\lambda\ge 0}$ be the bias-variance optimizing class of sensitivities that traces out the solutions to equation (10) as we vary the bound $\overline{B}$ when $\mathcal{C} = \mathcal{C}(1)$. The index λ determines the relative weight on the bias; it can correspond to the Lagrange multiplier in a Lagrangian formulation of (10), or we can simply take $\lambda = \overline{B}$ if we are solving (10) directly. Let $\overline{B}_\lambda = \overline{\text{bias}}_{\mathcal{C}(1)}(k_\lambda)$. It then follows by a change-of-variables argument that $\overline{\text{bias}}_{\mathcal{C}(M)}(k_\lambda) = M\overline{B}_\lambda$, and that k_λ minimizes the asymptotic variance subject to this bound on worst-case bias

over $\mathcal{C}(M)$. Thus, $\{k_\lambda\}_{\lambda\geq 0}$ is also a bias-variance optimizing class of sensitivities for $\mathcal{C}(M)$. We therefore only need to compute the class $\{k_\lambda\}_{\lambda\geq 0}$ only once, even when a range of values M is considered.

With this simplification, we can construct CIs for a range of values of M as follows:

1. Obtain an initial estimate $\hat{\theta}_{\text{initial}}$ and estimates $\hat{H}$, $\hat{\Gamma}$ and $\hat{\Sigma}$ of H, Γ, and Σ.

   In particular, for the GMM model, when $\hat{g}(\theta) = \frac{1}{n}\sum_{i=1}^n g(w_i,\theta)$, we can take $\hat{\theta}_{\text{initial}}$ to be the GMM estimator $\hat{\theta}_W = \operatorname{argmin}_\theta \hat{g}(\theta)' W \hat{g}(\theta)$ for some weight matrix W. The remaining objects are the usual quantities used to estimate the asymptotic variance of this estimator: $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n g(w_i,\hat{\theta}_{\text{initial}}) g(w_i,\hat{\theta}_{\text{initial}})'$ (or, in the case of dependent observations, an autocorrelation robust version of this estimate), $\hat{\Gamma} = \frac{d}{d\theta'}\hat{g}(\theta)|_{\theta=\hat{\theta}_{\text{initial}}}$ (or, if $g(\theta,w)$ is nonsmooth, a numerical derivative as in Hong, Mahajan, and Nekipelov 2015, or Section 7.3 of Newey and McFadden 1994) and $\hat{H} = \frac{d}{d\theta} h(\theta)|_{\theta=\hat{\theta}_{\text{initial}}}$.

2. Compute the bias-variance optimizing class $\{\hat{k}_\lambda\}_{\lambda\geq 0}$ that solves (10) with $\mathcal{C} = \mathcal{C}(1)$ and with $\hat{\Sigma}$ in place of Σ. Algorithms and closed-form solutions for computing $\{\hat{k}_\lambda\}_{\lambda\geq 0}$ for particular choices of $\mathcal{C}(M)$ are discussed in Remark 3.1. Let $\overline{B}_\lambda = \sup_{c\in\mathcal{C}(1)} |\hat{k}_\lambda' c|$. For each M, let λ_M^* minimize the CI length[4] $2\operatorname{cv}_\alpha(M\overline{B}_\lambda/\sqrt{\hat{k}_\lambda'\hat{\Sigma}\hat{k}_\lambda}) \cdot \sqrt{\hat{k}_\lambda'\hat{\Sigma}\hat{k}_\lambda}$ over λ.

3. For each M, construct the one-step estimator $\hat{h}_{\lambda_M^*} = h(\hat{\theta}_{\text{initial}}) + \hat{k}_{\lambda_M^*}'\hat{g}(\hat{\theta}_{\text{initial}})$, and report the misspecification-robust CI under $\mathcal{C}(M)$

$$\hat{h}_{\lambda_M^*} \pm \operatorname{cv}_\alpha\Big(M\overline{B}_{\lambda_M^*}/\sqrt{\hat{k}_{\lambda_M^*}'\hat{\Sigma}\hat{k}_{\lambda_M^*}}\Big)\cdot\sqrt{\hat{k}_{\lambda_M^*}'\hat{\Sigma}\hat{k}_{\lambda^*}/n}. \tag{11}$$

Remark 3.1 (Choice of $\mathcal{C}(M)$). A simple and flexible way of forming the set $\mathcal{C}$ is to take

$$\mathcal{C} = \mathcal{C}(M) = \big\{B\gamma\colon \|\gamma\| \leq M\big\}, \tag{12}$$

where B is a $d_g \times d_\gamma$ matrix and $\|\cdot\|$ is some norm. The matrix B can be used to standardize moments, account for their correlations, or to pick out which moments are believed to be misspecified. For instance, setting B to the last d_γ columns of the $d_g \times d_g$ identity matrix allows for misspecification in the last d_γ moments, while maintaining that the first $d_g - d_\gamma$ moments are valid.

In light of our result in Section 4 that it is not possible to determine the set $\mathcal{C}$ in a data-driven way, the normalizing matrix B and the baseline misspecification magnitude M used should be chosen to reflect application-specific arguments about which forms of misspecification are plausible; we can then vary M over other plausible choices as a form of sensitivity analysis. We illustrate this in the context of our application in Section 6, and we refer the reader to Conley, Hansen, and Rossi (2012) and Andrews, Gentzkow, and Shapiro (2017) for additional examples and discussion. Alternatively, one can also use measures of statistical distance such as the probability of detecting that the model is

[4]The critical value $\operatorname{cv}_\alpha(b)$ can easily be computed in statistical software as the square root of the $1-\alpha$ quantile of a noncentral χ^2 distribution with 1 degree of freedom and noncentrality parameter b^2.

misspecified to aid with interpretation of M, as suggested in Hansen and Sargent (2008) or Bonhomme and Weidner (2020).

While it is not possible to determine M automatically, it is possible to obtain a lower CI $[M_{\min}, \infty]$ for M, which can be used as a diagnostic check verifying that the values of M considered are not too small. We develop such tests by generalizing the J-test of overidentifying restrictions in Appendix B. We recommend reporting the lower bound $M_{\min}$ along with the plot of the optimal CI as a function of M. Reporting such a lower bound is in line with a recent proposal by Masten and Poirier (2020) to report results that are robust under a set $\mathcal{C}$ that is consistent with the observed data while being as small as possible in some sense.

The norm $\|\cdot\|$ determines how the researcher's bounds on each element of γ interact. With the ℓ_∞ norm, one places separate bounds on each element of γ, which leads to a simple interpretation: no single element of γ can be greater than M. Under an ℓ_p norm with $1 \leq p < \infty$, the bounds on each element of γ interact with each other, so that larger amounts of misspecification in one element is allowed if other elements are correctly specified. Depending on whether such interactions are desirable, we recommend setting $p = 2$, or $p = \infty$.

For these choices of the norm, computing the class of optimal sensitivities $\{\hat{k}_\lambda\}_{\lambda \geq 0}$ is particularly simple. In particular, when $\|\cdot\|$ corresponds to an ℓ_p norm, the worst-case bias has a closed form, since by Hölder's inequality, $\overline{\text{bias}}_{\mathcal{C}(M)}(k) = \sup_{\|\gamma\|_p \leq 1} M|k'B\gamma| = M\|B'k\|_{p'}$, where p' is the Hölder complement of p ($p' = 1$ if $p = \infty$, while $p' = 2$ if $p = 2$), and the optimal sensitivities $\{\hat{k}_\lambda\}_{\lambda \geq 0}$ can be computed by casting the problem as a penalized regression problem. We explain the connection to penalized regression, and provide details in Appendix A.2.

When $p = 2$, so that $\|\cdot\|$ corresponds to the Euclidean norm, the problem is analogous to ridge regression, and the optimal sensitivities in Step 2 of the implementation take the form $\hat{k}'_\lambda = -H(\Gamma' W_\lambda \Gamma)^{-1}\Gamma' W_\lambda$, where $W_\lambda = (\lambda BB' + \hat{\Sigma})^{-1}$, with $\overline{B}_\lambda = \|B'\hat{k}_\lambda\|_2$. As an alternative to using the one-step estimator in Step 3 of the implementation, one can implement this sensitivity directly as a GMM estimator with weighting matrix W_λ (see also Remark 3.2 below). Relative to the optimal weighting matrix Σ^{-1} under correct specification, the matrix W_λ trades off precision of the moments against their potential misspecification.

When $p = \infty$, the penalized regression analogy leads a simple algorithm for computing the optimal sensitivities $\{\hat{k}_\lambda\}_{\lambda \geq 0}$ that is similar to the LASSO/LAR algorithm (Efron et al. (2004)). We give details on the algorithm in Appendix A.2. It follows from this algorithm if B corresponds to columns of the identity matrix, as M grows, the optimal sensitivity successively drops the "least informative" moments, so that in the limit, if $d_g \leq d_\gamma + d_\theta$, the optimal sensitivity corresponds to that of an exactly identified GMM estimator based on the d_θ "most informative" moments only, where "informativeness" is given by both the variability of a given moment, and its potential misspecification. If $d_g > d_\gamma + d_\theta$, one simply drops all invalid moments in the limit.

REMARK 3.2. In Step 3 of our implementation, we use a one-step estimator $\hat{h}_\lambda$ to compute a CI that is asymptotically valid and optimal. Due to concerns about finite-sample

behavior (analogous to concerns about finite sample behavior of one-step estimators in the correctly specified case), one may prefer using a different estimator that is asymptotically equivalent to $\hat{h}_\lambda$. In general, one can implement an estimator with sensitivity k as a GMM or minimum distance estimator by using an appropriate weighting matrix, so that one can in particular replace $\hat{h}_\lambda$ by $h(\hat{\theta}_W)$, with the weighting matrix W appropriately chosen. To give the formula for the weighting matrix, let $\Gamma_\perp$ denote a $d_g \times (d_g - d_\theta)$ matrix that's orthogonal to Γ, so that $\Gamma_\perp'\Gamma = 0$, and let $\hat{\Gamma}_\perp$ denote a consistent estimate. Let S denote a $d_g \times d_\theta$ matrix that satisfies $S'\hat{\Gamma} = -I$ and $\hat{k}_\lambda = S\hat{H}'$. Then we can set $W = SW_1S' + \hat{\Gamma}_\perp W_2 \hat{\Gamma}_\perp'$ for some nonsingular matrix W_1, and an arbitrary conformable matrix W_2. It can be verified by simple algebra that $\hat{\theta}_W$ will have sensitivity k_λ.

Remark 3.3 (Global misspecification). While we focus on the local misspecification setting, in which the set $\mathcal{C}/\sqrt{n}$ shrinks with n at a $1/\sqrt{n}$, one can use our insights about optimal weighting to construct a CI that retains the near-optimality properties of the above CI under local misspecification, while having correct coverage under asymptotics in which this set shrinks more slowly or stays fixed with the sample size (the latter is termed "global misspecification" in the literature). Let W be a weighting matrix that leads to the optimal sensitivity, as described in Remark 3.2 above, and let $\mathcal{I}_{\tilde{c}}$ be a CI constructed from the GMM estimator with moment conditions $\theta \mapsto g(w_i, \theta) - \tilde{c}$. Let $\mathcal{I} = \bigcup_{\tilde{c} \in \mathcal{C}/\sqrt{n}} \mathcal{I}_{\tilde{c}}$ be the union of these CIs over possible values of $\tilde{c}$ in the set $\mathcal{C}/\sqrt{n}$. Such an approach was suggested in the context of misspecified linear IV by Conley, Hansen, and Rossi (2012), although they did not consider adjusting the weighting matrix. The resulting CI has correct asymptotic coverage under both local and global misspecification, and, for one-sided CI construction, is asymptotically equivalent under local misspecification to the CI discussed above. We provide further details in Appendix D. In the Appendix, we also discuss a second approach to constructing CIs valid under global misspecification based on misspecification-robust standard errors (Hall and Inoue (2003)), which is applicable if the estimate of the worst-case bias under global misspecification is asymptotically normal. The resulting one- and two-sided CIs are asymptotically equivalent under local misspecification to the optimal CIs discussed above.

Remark 3.4 (Other performance criteria). In addition to constructing a CI, one may be interested in a point estimate of $h(\theta_0)$, using mean squared error (MSE) as the criterion. The steps to forming the MSE optimal point estimate are exactly the same as above, except that, rather than minimizing CI length in Step 2, we choose λ to minimize $\overline{\text{bias}}_{\mathcal{C}}(\hat{k}_\lambda)^2 + \hat{k}_\lambda'\hat{\Sigma}\hat{k}_\lambda = M\overline{B}_\lambda + \hat{k}_\lambda'\hat{\Sigma}\hat{k}_\lambda$. Similar ideas apply to other criteria, such as mean absolute deviation or quantiles of excess length of one-sided CIs (discussed in Appendix C). If λ is chosen differently in Step 2 the CI computed in Step 3 will be longer than the one computed at λ_M^*, but it will still have correct coverage.

4. Efficiency bounds and near optimality

The CI given in equation (11) has the apparent defect that the local misspecification vector c is reflected in the length of the CI only through the a priori restriction $\mathcal{C}$ imposed

by the researcher. Thus, if the researcher is conservative about misspecification, the CI will be wide, even if it turns out that c is in fact much smaller than the a priori bounds defined by $\mathcal{C}$. Moreover, this approach requires the researcher to explicitly specify the set $\mathcal{C}$, including any tuning parameters such as the parameter M if the set takes the form $\mathcal{C} = \mathcal{C}(M)$ that we considered in Section 3. One may therefore seek to improve upon this CI by estimating the magnitude of c, or by estimating the tuning parameters, and constructing a CI that is shorter if these estimates indicate that misspecification is mild. Similarly, it may be restrictive to require that the CI be centered at an asymptotically linear estimator: this rules out, for example, using a J-test to decide which moments to use.

The main result of this section shows that, when $\mathcal{C}$ is convex and centrosymmetric ($c \in \mathcal{C}$ implies $-c \in \mathcal{C}$), the scope for improving on the CI in (11) is nonetheless limited: no sequence of CIs that maintain coverage under all local misspecification vectors $c \in \mathcal{C}$ can be substantially tighter, even under correct specification. This result can be interpreted as translating results from a "limiting experiment" that is an extension of the linear regression model. We first give a heuristic derivation of this limiting experiment and explain our result in the context of this limiting experiment. We then present the formal asymptotic result, and discuss its implications in some familiar settings. Readers who are interested only in implementing the methods, rather than efficiency results, can skip this section.

We restrict attention in this section to the GMM model, in which $\hat{g}(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(w_i, \theta)$, and we further restrict the data $\{w_i\}_{i=1}^{n}$ to be independent and identically distributed (i.i.d.). Similar to semiparametric efficiency theory in the standard, correctly specified case, this facilitates parts of the formal statements and proofs, such as the definition of the set of distributions under which coverage is required and the construction of least favorable submodels. We expect that analogous results could be obtained in other settings.

4.1 *Limiting experiment*

As discussed in Section 2.1, we can form CIs based on linear estimators with asymptotic distribution $\mathcal{N}(k'c, k'\Sigma k)$. This suggests that the problem of constructing an asymptotically valid CI for $h(\theta)$ in the model (1) is asymptotically equivalent to the problem of constructing a CI for the parameter $H\theta$ in the approximately linear model

$$Y = -\Gamma\theta + c + \Sigma^{1/2}\varepsilon, \quad c \in \mathcal{C}, \varepsilon \sim \mathcal{N}(0, I), \tag{13}$$

where Γ, H, and $\Sigma^{1/2}$ are known, and we observe Y. One can think of this model as an "approximately" linear regression model, with $-\Gamma$ playing the role of the design matrix of the (fixed) regressors, and c giving the approximation error. This model dates back at least to Sacks and Ylvisaker (1978), who considered estimation in this model when $\mathcal{C}$ is a rectangular set and Σ is diagonal. The analog of the asymptotically linear estimator $\hat{h}$ in (4) is the linear estimator $k'Y$. To see the analogy, note that $k'Y - H\theta$ is distributed $\mathcal{N}((-k'\Gamma - H)\theta + k'c, k'\Sigma k)$, and restricting ourselves to estimators that do not have

infinite worst-case bias when θ is unrestricted gives the condition (5). In the limiting experiment, the analog of the CI (7) is given by the CI $k'Y \pm \mathrm{cv}_\alpha(\overline{\mathrm{bias}}_{\mathcal{C}}(k)/\sqrt{k'\Sigma k}) \cdot \sqrt{k'\Sigma k}$. Finding weights that minimize the length of this CI is isomorphic to the problem of finding the sensitivity that minimizes the asymptotic CI length in (9).

For a general convex set $\mathcal{C}$, the bias-variance optimization problem in equation (10) can be reformulated as a convex programming problem, as shown in Low (1995). In particular, when the set $\mathcal{C}$ is centrosymmetric (see Appendix C for the general case), the bias-variance optimizing class of weights $\{k_\lambda\}_{\lambda>0}$ is given by the class $\{k_\delta\}_{\delta>0}$, where

$$k'_\delta = k'_{\delta,\Sigma,\Gamma,H,\mathcal{C}} = \frac{-(c_\delta - \Gamma\theta_\delta)'\Sigma^{-1}}{(c_\delta - \Gamma\theta_\delta)'\Sigma^{-1}\Gamma H'/HH'}, \tag{14}$$

and, for each δ, c_δ, θ_δ are the solutions to the convex program

$$\sup_{\theta,c} H\theta \quad \text{s.t. } c \in \mathcal{C}, (c-\Gamma\theta)'\Sigma^{-1}(c-\Gamma\theta) \le \delta^2/4. \tag{15}$$

It then follows from Donoho (1994) that among fixed-length CIs based on linear estimators (CIs that take the form $k'Y \pm \chi$ for some constant χ), the shortest CI in the limiting experiment takes the form

$$k'_{\delta^*}Y \pm \mathrm{cv}_\alpha\left(\overline{\mathrm{bias}}_{\mathcal{C}}(k_{\delta^*})/\sqrt{k'_{\delta^*}\hat{\Sigma}k_{\delta^*}}\right) \cdot \sqrt{k'_{\delta^*}\Sigma k_{\delta^*}}, \tag{16}$$

where $\overline{\mathrm{bias}}_{\mathcal{C}}(k_\delta) = -k'_\delta c_\delta$, and $\delta^* = \mathrm{argmin}_{\delta>0} 2\,\mathrm{cv}_\alpha(\overline{\mathrm{bias}}_{\mathcal{C}}(k_\delta)/\sqrt{k'_\delta\Sigma k_\delta}) \cdot \sqrt{k'_\delta\Sigma k_\delta}$ is chosen to minimize the CI length. The CI in (11) is an analog of this CI, with δ playing the role of the index λ.

The CI in (16) takes a familiar form in the special case in which $\mathcal{C}$ is a linear subspace of $\mathbb{R}^{d_g}$, so that for some $d_g \times d_\gamma$ full-rank matrix B with $d_\gamma \le d_g - d_\theta$, $\mathcal{C} = \{B\gamma\colon \gamma \in \mathbb{R}^{d_\gamma}\}$. Let $B_\perp$ denote a $d_g \times (d_g - d_\gamma)$ matrix that's orthogonal to B. Then for any $\delta > 0$, $k'_\delta = k'_{\mathrm{LS},B}$, where

$$k'_{\mathrm{LS},B} = -H\left(\Gamma'B_\perp\left(B'_\perp\Sigma B_\perp\right)^{-1}B'_\perp\Gamma\right)^{-1}\Gamma'B_\perp\left(B'_\perp\Sigma B_\perp\right)^{-1}B'_\perp \tag{17}$$

is the sensitivity of the GLS estimator after premultiplying (13) by $B'_\perp$, (which effectively picks out the observations with zero misspecification). Since this estimator is unbiased, the CI in (16) becomes $k'_{\mathrm{LS},B}Y \pm z_{1-\alpha/2}\sqrt{k'_{\mathrm{LS},B}\Sigma k_{\mathrm{LS},B}}$.

Like the asymptotic CI (11), the CI in (16) has the potential drawback that its length is determined by the worst possible misspecification in $\mathcal{C}$, leaving open the possibility of efficiency improvements when c turns out to be close to zero. As a best-case scenario for such improvements, consider the problem: among confidence sets with coverage at least $1-\alpha$ for all $\theta \in \mathbb{R}^{d_\theta}$ and $c \in \mathcal{C}$, minimize expected length when $\theta = \theta^*$ and $c = 0$. Note that this setup is even more favorable for potential improvements on our CI, since it allows the researcher to guess correctly that θ is equal to some θ^*, and it allows for confidence sets that are not intervals (in this case, length is defined as Lebesgue measure). Let $\kappa_*(H,\Gamma,\Sigma,\mathcal{C})$ denote the ratio of this optimized expected length relative to the length of the CI in (16) (it can be shown that this ratio does not depend on θ^*).

If $\mathcal{C}$ is convex, a formula for $\kappa_*(H,\Gamma,\Sigma,\mathcal{C})$ follows from applying the general results in Corollary 3.3 in Armstrong and Kolesár (2018) to the limiting model. If $\mathcal{C}$ is also centrosymmetric, then

$$\kappa_*(H,\Gamma,\Sigma,\mathcal{C}) = \frac{(1-\alpha)E\big[\omega\big(2(z_{1-\alpha}-Z)\big) \mid Z \le z_{1-\alpha}\big]}{2\min_\delta \mathrm{cv}_\alpha\left(\frac{\omega(\delta)}{2\omega'(\delta)} - \frac{\delta}{2}\right)\omega'(\delta)}, \tag{18}$$

where $Z \sim \mathcal{N}(0,1)$ and $\omega(\delta)$ is two times the optimized value of (15). Furthermore, we show in Theorem C.7 that the right-hand side is lower-bounded by $(z_{1-\alpha}(1-\alpha) - \tilde{z}_\alpha\Phi(\tilde{z}_\alpha)+\phi(z_{1-\alpha})-\phi(\tilde{z}_\alpha))/z_{1-\alpha/2}$, where $\tilde{z}_\alpha = z_{1-\alpha}-z_{1-\alpha/2}$ for any H,Γ,Σ, and $\mathcal{C}$, where $\phi(\cdot)$ denotes the standard normal density. For $\alpha = 0.05$, this universal lower bound evaluates to 71.7%. Evaluating κ_* for particular choices of H, Γ, Σ, and $\mathcal{C}$ often yields even higher efficiency.

If $\mathcal{C}$ is a linear subspace, then $\omega(\delta)$ is linear, and

$$\kappa_*(H,\Gamma,\Sigma,\mathcal{C}) = \frac{(1-\alpha)z_{1-\alpha}+\phi(z_{1-\alpha})}{z_{1-\alpha/2}} \ge \frac{z_{1-\alpha}}{z_{1-\alpha/2}}, \tag{19}$$

where the lower bound follows since $\phi(z_{1-\alpha}) \ge \alpha z_{1-\alpha}$ by the Gaussian tail bound $1-\Phi(x) \le \phi(x)/x$ for $x>0$. This bound corresponds to that in Pratt (1961) for the case of a univariate normal mean. The potential efficiency improvement essentially comes from using prior knowledge of θ^* to turn a two-sided critical value into a one-sided critical value. Furthermore, it follows from Joshi (1969) that the CI $k'_{\mathrm{LS},B}Y \pm z_{1-\alpha/2}\sqrt{k'_{\mathrm{LS},B}\Sigma k_{\mathrm{LS},B}}$ is the unique CI that achieves minimax expected length. Thus, not only is the scope for improvement at a particular θ^* bounded by (19), any CI with shorter expected length at some θ^* must necessarily perform worse elsewhere in the parameter space.

For the one-sided CI (8), the analogous CI in the limiting experiment is $[k'Y - \overline{\mathrm{bias}}_{\mathcal{C}}(k) - z_{1-\alpha}\sqrt{k'\Sigma k}, \infty)$, and, as we discuss in Appendix C, to choose the optimal sensitivity k, one can consider optimizing a given quantile of its worst-case excess length. The results in Armstrong and Kolesár (2018) again give an efficiency bound for improvement at $c=0$ and a particular θ^*, analogous to (18) for the two-sided case. See Appendix C for details. If $\mathcal{C}$ is a linear subspace, then optimizing quantiles of worst-case excess length yields the CI $[k'_{\mathrm{LS},B}Y - z_{1-\alpha}\sqrt{k'_{\mathrm{LS},B}\Sigma k_{\mathrm{LS},B}}, \infty)$, independently of the quantile one is optimizing. Furthermore, the efficiency bound implies that this one-sided CI is in fact fully optimal over all quantiles of excess length and all values of θ, c in the local parameter space.

These efficiency results for the CI (16) in the limiting experiment suggest that the scope for improvement over the CI in (11) should be limited in large samples. Theorem 4.1, stated in the next section, uses the analogy with the approximately linear model (13) along with Le Cam-style arguments involving least favorable submodels to show that this bound indeed translates to the locally misspecified GMM model. For one-sided CIs, we state an analogous result in Appendix C. We discuss the implications of these results in Section 4.3.

4.2 *Asymptotic efficiency bound*

To make precise our statements about coverage and efficiency, we need the notion of uniform (in the underlying distribution) coverage of a confidence interval. This requires additional notation, which we now introduce. Let $\mathcal{P}$ denote a set of distributions P of the data $\{w_i\}_{i=1}^n$, and let $\Theta_n \subseteq \mathbb{R}^{d_\theta}$ denote the parameter space for θ. We require coverage for all pairs $(\theta, P) \in \Theta_n \times \mathcal{P}$ such that $\sqrt{n} g_P(\theta) \in \mathcal{C}$, where the subscript P on the population moment condition makes it explicit that it depends on the distribution of the data.[5] Letting $\mathcal{S}_n = \{(\theta, P) \in \Theta_n \times \mathcal{P} : \sqrt{n} g_P(\theta) \in \mathcal{C}\}$ denote this set, the condition for coverage at confidence level $1-\alpha$ can be written

$$\liminf_{n\to\infty} \inf_{(\theta,P)\in\mathcal{S}_n} P\big(h(\theta) \in \mathcal{I}_n\big) \geq 1-\alpha. \tag{20}$$

We say that a confidence set $\mathcal{I}_n$ is asymptotically valid (uniformly over $\mathcal{S}_n$) at confidence level $1-\alpha$ if this condition holds.[6]

Among two-sided CIs of the form $\hat{h} \pm \hat{\chi}$ that are asymptotically valid, we prefer CIs with shorter expected length. To avoid issues with convergence of moments, we use truncated expected length, and define the asymptotic expected length of a two-sided CI at $P_n \in \mathcal{P}$ as $\liminf_{T\to\infty} \liminf_{n\to\infty} E_{P_n} \min\{\sqrt{n} \cdot 2\hat{\chi}, T\}$, where E_P denotes expectation under P.

We are now ready to state the main efficiency result.

THEOREM 4.1. *Suppose that $\mathcal{C}$ is convex and centrosymmetric. Let $\hat{h}_{\lambda^*}$ and $\hat{\chi}^*_{\lambda^*}$ be formed as in Section 3. Suppose that Assumptions C.2, C.3, C.5, C.6, and C.7 in Appendix C hold. Suppose that the data $\{w_i\}_{i=1}^n$ are i.i.d. under all $P \in \mathcal{P}$. Let (θ^*, P_0) be correctly specified (i.e., $g_{P_0}(\theta^*) = 0$) such that $\mathcal{P}$ contains a submodel through P_0 satisfying Assumption C.1. Then*:

(i) *The CI $\hat{h}_{\lambda^*} \pm \hat{\chi}^*_{\lambda^*}$ is asymptotically valid, and its half-length $\hat{\chi}^*_{\lambda^*}$ satisfies $\sqrt{n}\hat{\chi}^*_{\lambda^*} = \chi(\theta, P) + o_P(1)$ uniformly over $(\theta, P) \in \mathcal{S}_n$ where*

$$\chi(\theta, P) = \min_k \mathrm{cv}_\alpha\big(\overline{\mathrm{bias}}_{\mathcal{C}}(k)/\sqrt{k'\Sigma_{\theta,P}k}\big)\sqrt{k'\Sigma_{\theta,P}k}$$

with $\overline{\mathrm{bias}}_{\mathcal{C}}(k)$ calculated with $\Gamma = \Gamma_{\theta,P}$ and $H = H_\theta$.

(ii) *For any other asymptotically valid CI $\hat{h} \pm \hat{\chi}$,*

$$\frac{\liminf_{T\to\infty} \liminf_{n\to\infty} E_{P_0} \min\{\sqrt{n} \cdot 2\hat{\chi}, T\}}{2\chi\big(\theta^*, P_0\big)} \geq \kappa_*(H_{\theta^*}, \Gamma_{\theta^*,P_0}, \Sigma_{\theta^*,P_0}, \mathcal{C}),$$

[5] To be precise, we should also subscript all other quantities such as Γ and Σ by P. To prevent notational clutter, we drop this index in the main text unless it causes confusion.

[6] In general, θ_0 and $h(\theta_0)$ may be set identified for a given sample size n (although our assumptions imply that the identified set will shrink at a root-n rate). The coverage requirement (20) states that the CI must cover points in the identified set for $h(\theta)$, as in Imbens and Manski (2004); see Appendix C.

where $\kappa_*(H,\Gamma,\Sigma,\mathcal{C})$ is defined in (18). Furthermore, for any H, Σ, Γ, and $\mathcal{C}$, κ_* admits the universal lower bound $(z_{1-\alpha}(1-\alpha)-\tilde{z}_\alpha\Phi(\tilde{z}_\alpha)+\phi(z_{1-\alpha})-\phi(\tilde{z}_\alpha))/z_{1-\alpha/2}$, where $\tilde{z}_\alpha = z_{1-\alpha}-z_{1-\alpha/2}$ and $\phi(\cdot)$ denotes the standard normal density.

The proof for this theorem is given in Appendix C, which also gives an analogous result for one-sided confidence intervals. Assumptions C.2, C.3, C.5, C.6, and C.7, stated in Appendix C, require that the conditions in Section 2 hold in a uniform sense over the class $\mathcal{P}$. In the Supplemental Materials, we give primitive conditions for these assumptions in the misspecified linear IV model. Assumption C.1, also stated in the Appendix, requires that the class $\mathcal{P}$ be rich enough to contain a submodel that is least favorable for the GMM problem, so that the class does not implicitly impose any other conditions that could be used to make inference easier. In the Supplemental Materials, we provide a general way of constructing a submodel satisfying these conditions.

The universal lower bound on κ_* is new and may be of independent interest. For $\alpha = 0.05$, it evaluates to 71.7%. The universal lower bound is sharp in the sense that there exist Γ, Σ, H, and $\mathcal{C}$ for which κ_* equals this lower bound. In particular applications, the efficiency bound κ_* can be computed at estimates of Γ, Σ and H, and often, this gives much higher efficiencies. We illustrate these bounds in the empirical application in Section 6.

4.3 *Discussion*

To help build intuition for the efficiency bound in Theorem 4.1, and to relate this result to the literature, we now consider some special cases. We first discuss the (standard) correctly specified case. Second, we consider the case in which some moments are known to be valid, and the misspecification in the remaining moments is unrestricted. This case may be of interest in its own right. We then discuss the general case. Finally, we discuss the connection to certain statistical measures of distance considered in the literature.

4.3.1 *Correctly specified case* Suppose that $\mathcal{C} = \{0\}$. This is in particular a linear subspace of $\mathbb{R}^{d_g}$, with $B = 0$, and $B_\perp = I$, the $d_g \times d_g$ identity matrix. Thus, in the limiting experiment, the optimal CI uses the GLS estimator $k'_{\mathrm{LS},0}Y$, with $k_{\mathrm{LS},0}$ given in (17) (with $B = 0$). For testing the null hypothesis $H\theta = h_0$ against the one-sided alternative $H\theta \geq h_0$, the one-sided z-statistic based on $k'_{\mathrm{LS},0}Y$ is uniformly most powerful (van der Vaart (1998, Proposition 15.2)). Inverting these tests yields the CI $[k'_{\mathrm{LS},0}Y - z_{1-\alpha}\sqrt{k'_{\mathrm{LS},0}\Sigma k_{\mathrm{LS},0}}, \infty)$. Since the underlying tests are uniformly most powerful, this CI achieves the shortest excess length, simultaneously for all quantiles and all possible values of the parameter θ. For two-sided CIs, the results described in Section 4.1 imply that the CI $h'_{\mathrm{LS},0}Y \pm z_{1-\alpha/2}\sqrt{k'_{\mathrm{LS},0}\Sigma k_{\mathrm{LS},0}}$ is the unique CI that achieves minimax expected length, and the efficiency of this CI relative to a CI that optimizes its expected length at a single value θ^* of θ when indeed $\theta = \theta^*$ is given in equation (19). It evaluates to 84.99% at $\alpha = 0.05$.

Applying Theorem 4.1 to the case $\mathcal{C} = \{0\}$ gives an asymptotic version of the two-sided efficiency bound. Furthermore, the CI in Theorem 4.1 reduces to the usual two-sided CI based on $\hat{\theta}_{\Sigma^{-1}}$. Thus, in this case, Theorem 4.1 shows that very little can be

gained over the usual two-sided CI by optimizing the CI relative to a particular distribution P_0. Results in the appendix give an analogous result for one-sided CIs. In the one-sided case, this asymptotic result is essentially a version of a classic result from the semiparametric efficiency literature for one-sided tests, applied to CIs (see Chapter 25.6 in van der Vaart (1998)). In the two-sided case, the result is, to our knowledge, new.

4.3.2 *Some valid and some invalid moments* Consider now the case in which the first $d_g - d_\gamma$ moments are known to be valid, with the potential misspecification for the remaining d_γ moments unrestricted. Then $\mathcal{C} = \{(0', \gamma')' : \gamma \in \mathbb{R}^{d_\gamma}\}$ corresponds to a linear subspace with B given by the last d_γ columns of the identity matrix, and $B_\perp$ given by the first $d_g - d_\gamma$ columns. Optimal CIs in the limiting experiment therefore use the estimator $k'_{\mathrm{LS},B}Y$, which is the GLS estimator based only on the observations with no misspecification.

The one-sided CI based on $k'_{\mathrm{LS},B}Y$ achieves the shortest excess length, simultaneously for all quantiles and all possible values of the parameter θ. The two-sided CI $k'_{\mathrm{LS},B}Y \pm z_{1-\alpha/2}\sqrt{k'_{\mathrm{LS},B}\Sigma k_{\mathrm{LS},B}}$ is optimal in the same sense as the usual CI in Section 4.3.1: it achieves minimax expected length, and its efficiency, relative to a CI that optimizes its length at a single θ^* and $\gamma = 0$, is lower-bounded by $z_{1-\alpha}/z_{1-\alpha/2}$. Theorem 4.1 formally translates the efficiency bound from the limiting model to the GMM model, so that the usual two-sided CI based on $h(\hat{\theta}_{W(B)})$ is asymptotically efficient in the same sense as the usual CI based on $h(\hat{\theta}_{\Sigma^{-1}})$ discussed in Section 4.3.1 under correct specification. Just as with the results in Section 4.3.1, this asymptotic result is, to our knowledge, new. The one-sided analog follows from the results in Appendix C. These results stand in sharp contrast to the results for estimation, where the MSE improvement at small values of γ may be substantial.

An important consequence of these results is that asymptotically valid one-sided CIs based on shrinkage or model-selection procedures, such as one-sided versions of the CIs proposed in Andrews and Guggenberger (2009), DiTraglia (2016) or McCloskey (2020) must have *worse* excess length performance than the usual one-sided CI based on the GMM estimator $h(\hat{\theta}_{W(B)})$ that uses valid moments only. While it is possible to construct two-sided CIs that improve upon the usual CI based on $h(\hat{\theta}_{W(B)})$ at particular values of θ and γ, the scope for such improvement is smaller than the ratio of one- to two-sided critical values. Furthermore, any such improvement must come at the expense of worse performance at other points in the parameter space.[7] Therefore, in order to tighten CIs based on valid moments only, it is *necessary* to make a priori restrictions on the potential misspecification of the remaining moments.

4.3.3 *General case* According to the results in Section 4.3.2, one must place a priori bounds on the amount of misspecification in order to use misspecified moments. This leads us to the general case, where we place the local misspecification vector c in some set $\mathcal{C}$ that is not necessarily a linear subspace. One can then form a CI centered at an estimate formed from these misspecified moments using the methods in Section 3. In the

[7] Consistently with these results, in a simulation study considered in DiTraglia (2016), the post-model selection CI that he proposes is shown to be wider on average than the usual CI around a GMM estimator that uses valid moments only.

case where $\mathcal{C}$ is convex and centrosymmetric, Theorem 4.1 shows that this CI is near optimal, in the sense that no other CI can improve upon it by more than a factor of κ_*, even in the favorable case of correct specification. Since the width of the CI is asymptotically constant under local parameter sequences $\theta_n \to \theta^*$ and sufficiently regular probability distributions $P_n \to P_0$ (e.g., $P_n \to P_0$ along submodels satisfying Assumption C.1), this also shows that the CI is near optimal in a local minimax sense. In the general case, Theorem 4.1, as well as the analogous results for one-sided CIs in Appendix C are, to our knowledge, new.

As we discuss in Section 3, we recommend reporting results for a range of sets $\mathcal{C}(M)$ indexed by a scalar M that bounds the magnitude of misspecification. One may instead wish to report a single CI based on a data-driven estimate of M, for example, by using a first-stage J test to assess plausible magnitudes of misspecification. Formally, one would seek a CI that is valid over $\mathcal{C}(\overline{M})$ while improving length when in fact $\|\gamma\| \ll \overline{M}$, where $\overline{M}$ is some initial conservative bound. When $\mathcal{C}$ is convex and centrosymmetric, Theorem 4.1 shows that the scope for such improvements is limited: the average length of any such CI cannot be much smaller than the CI that uses the most conservative choice $\overline{M}$, even when $c = 0$. The impossibility of choosing M based on the data is related to the impossibility of using specification tests to form an upper bound for M. On the other hand, it is possible to obtain a lower bound for M using such tests. We develop lower CIs for M in Appendix B.

4.3.4 *Cressie–Read divergences* Andrews, Gentzkow, and Shapiro (2020) have shown that defining misspecification in terms of the magnitude of any divergence in the Cressie and Read (1984) family leads to a set $\mathcal{C}$ that asymptotically takes the form $\mathcal{C} = \{\Sigma^{1/2}\gamma : \|\gamma\|_2 \le M\} = \{c : c'\Sigma^{-1}c \le M^2\}$. The Cressie–Read family includes the Hellinger distance used by Kitamura, Otsu, and Evdokimov (2013), who consider minimax point estimation among estimators satisfying certain regularity conditions. Since this set $\mathcal{C}$ takes the form discussed in Remark 3.1 with $p = 2$ and $B = \Sigma^{1/2}$, it follows from the discussion in Remark 3.1 that the optimal sensitivity corresponds to the GMM estimator with weighting matrix $(\lambda BB' + \Sigma)^{-1} = (\lambda + 1)^{-1}\Sigma^{-1}$. Since this is proportional to the weighting matrix Σ^{-1} that is optimal under correct specification, we obtain the same optimal sensitivity $k'_{\mathrm{LS},0} = -H(\Gamma'\Sigma^{-1}\Gamma)^{-1}\Gamma'\Sigma^{-1}$ as in the correctly specified case discussed in Section 4.3.1. As we show in Appendix A.1, this form of $\mathcal{C}$ leads to a closed form solution for the efficiency bound κ_*.

The results above imply that any estimator with sensitivity $k_{\mathrm{LS},0}$ is near optimal for CI construction. In line with these results, the estimator in Kitamura, Otsu, and Evdokimov (2013) has sensitivity $k_{\mathrm{LS},0}$. Thus, the usual GMM estimator $h(\hat{\theta}_{\Sigma^{-1}})$ and the estimator in Kitamura, Otsu, and Evdokimov (2013) are both near-optimal for CI construction, even if one allows for arbitrary CIs that are not necessarily centered at estimators that satisfy the regularity conditions in Kitamura, Otsu, and Evdokimov (2013). Also, because they have the same sensitivity, under this form of misspecification, the usual GMM estimator $h(\hat{\theta}_{\Sigma^{-1}})$ and the estimator in Kitamura, Otsu, and Evdokimov (2013) have the same local asymptotic minimax properties.

4.4 *Extensions: Asymmetric constraints and constraints on θ*

If the set $\mathcal{C}$ is convex but asymmetric (such as when $\mathcal{C}$ includes bounds on a norm as well as sign restrictions, or when $\mathcal{C}$ includes equality and sign restrictions, as in Moon and Schorfheide (2009)), one can still apply bounds from Armstrong and Kolesár (2018) to the limiting model described in Section 4.1. Our general asymptotic efficiency bounds in Appendix C translate these results to the locally misspecified GMM model so long as $\mathcal{C}$ is convex. Since the negative implications for efficiency improvements under correct specification use centrosymmetry of $\mathcal{C}$, introducing asymmetric restrictions, such as sign restrictions, is one possible way of getting efficiency improvements at some smaller set $\mathcal{D} \subseteq \mathcal{C}$ while maintaining coverage over $\mathcal{C}$. We derive efficiency bounds and optimal CIs for this problem in Appendix C. Interestingly, the scope for efficiency improvements can be different for one- and two-sided CIs, and can depend on the direction of the CI in this case. To get some intuition for this, note that, in the instrumental variables model with a single instrument and single endogenous regressor, sign restrictions on the covariance of an instrument with the error term can be used to sign the direction of the bias of the instrumental variables estimator, which is useful for forming a one-sided CI only in one direction.

Finally, while we focus on restrictions on c, one can also incorporate local restrictions on θ. Our general results in Appendix C give efficiency bounds that cover this case. Similar to the discussion above, these results have implications for using prior information about θ to determine the amount of misspecification, or to shrink the width of a CI directly. In particular, while it is possible to use prior information on θ (say, an upper bound on $\|\theta\|$ for some norm $\|\cdot\|$) to shrink the width of the CI, the width of the CI and the estimator around which it is centered must depend on the a priori upper bounds on the magnitude of θ and c when this prior information takes the form of a convex, centrosymmetric set for $(\theta', c')'$. This rules out, for example, choosing the moments based on whether the resulting estimate for θ is in a plausible range.

5. Applications

This section describes particular applications of our approach, along with a discussion of implementation details appropriate to each application.

5.1 *Instrumental variables*

The single equation linear instrumental variables (IV) model is given by

$$y_i = x_i'\theta_0 + \varepsilon_i, \tag{21}$$

where, in the correctly specified case, $E\varepsilon_i z_i = E(y_i - x_i'\theta_0)z_i = 0$, with z_i a d_g-vector of instruments. This is an instance of a GMM model with $g(\theta) = E(y_i - x_i'\theta)z_i$ and $\hat{g}(\theta) = \frac{1}{n}\sum_{i=1}^n z_i(y_i - x_i'\theta)$.

One common reason for misspecification in this model is that the instruments do not satisfy the exclusion restriction, because they appear directly in the structural equation (21), so that $\varepsilon_i = z_{Ii}'\gamma/\sqrt{n} + \eta_i$, where $E[z_i\eta_i] = 0$, and z_{Ii} corresponds to a subset I

of the instruments, the validity of which one is worried about. This form of misspecification has previously been considered in a number of papers, including Hahn and Hausman (2005), Conley, Hansen, and Rossi (2012), and Andrews, Gentzkow, and Shapiro (2017), among others. Bounding the norm of γ using some norm $\|\cdot\|$ then leads to the set given in equation (12), with $B = E[z_i z_{Ii}']$.

Although the matrix B is unknown, for the purposes of estimating the optimal sensitivity and constructing asymptotically valid CIs, it can be replaced by the sample analog $\hat{B} = n^{-1}\sum_{i=1}^n z_i z_{Ii}'$. This does not affect the asymptotic validity or coverage properties of the resulting CI. Under this setup, the parameter M bounds that magnitude of γ, the direct effect of the instruments on the outcome. Therefore, the appropriate choice of M will depend on the plausible magnitude of these direct effects; see, for example, Conley, Hansen, and Rossi (2012) for examples and a discussion.

The linearity of the moment condition leads to simplifications in our implementation in Section 3. In Step 1, as the initial estimator, one can use the two-stage least squares (2SLS) estimator

$$\hat{\theta}_{\text{initial}} = \left[\left(\sum_{i=1}^n z_i x_i'\right)'\left(\sum_{i=1}^n z_i z_i'\right)^{-1}\left(\sum_{i=1}^n z_i x_i'\right)\right]^{-1}\left(\sum_{i=1}^n z_i x_i'\right)'\left(\sum_{i=1}^n z_i z_i'\right)^{-1}\sum_{i=1}^n z_i y_i.$$

This leads to the estimates $\hat{\Gamma} = -\frac{1}{n}\sum_{i=1}^n z_i x_i'$ and $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n (y_i - x_i'\hat{\theta}_{\text{initial}})^2 z_i z_i'$. Alternatively, if we assume homoskedasticity, we can use the estimator $\hat{\Sigma}_H = \frac{1}{n}\sum_{i=1}^n (y_i - x_i'\hat{\theta}_{\text{initial}})^2 \cdot \frac{1}{n}\sum_{i=1}^n z_i z_i'$. In the correctly specified case, the 2SLS estimator is only optimal under homoskedasticity, while the GMM estimator with weighting matrix $\hat{\Sigma}^{-1}$ is optimal in general. Due to concerns with finite sample performance, however, it is common to use the 2SLS estimator along with standard errors based on a robust variance estimate, even when heteroskedasticity is suspected. Mirroring this practice, one can use $\hat{\Sigma}_H$ when forming the optimal sensitivity $\hat{k}_{\lambda_M^*}$ and worst-case bias in Step 2, but use the robust variance estimate $\hat{\Sigma}$ in Step 3 when forming the final CI in equation (11). The CI will then be optimal under homoskedasticity, but it will remain valid under heteroskedasticity, just like the usual CI based on 2SLS with robust standard errors in the correctly specified case.

If the parameter of interest linear in θ, $h(\theta) = H\theta$, then the one-step estimator $\hat{h}_{\lambda_M^*}$ in Step 3 does not depend on the choice of the initial estimator (except possibly through the estimate of Σ when forming the desired sensitivity):

$$\begin{aligned}\hat{h}_{\lambda_M^*} &= H\hat{\theta}_{\text{initial}} + \hat{k}_{\lambda_M^*}'\frac{1}{n}\sum_{i=1}^n (y_i - x_i'\hat{\theta}_{\text{initial}})z_i = \hat{k}_{\lambda_M^*}'\frac{1}{n}\sum_{i=1}^n y_i z_i + \left(H - \hat{k}_{\lambda_M^*}'\frac{1}{n}\sum_{i=1}^n z_i x_i'\right)\hat{\theta}_{\text{initial}}\\ &= \hat{k}_{\lambda_M^*}'\frac{1}{n}\sum_{i=1}^n y_i z_i,\end{aligned}$$

where the second line follows since the sensitivities $\hat{k}_{\lambda_M^*}$ satisfy $H = -\hat{k}_{\lambda_M^*}'\hat{\Gamma} = \hat{k}_{\lambda_M^*}'\frac{1}{n}\times\sum_{i=1}^n z_i x_i'$. Since the estimator $\hat{h}_{\lambda_M^*}$ is linear, the worst-case bias calculations are the same

under global misspecification, when the magnitude M of γ in equation (12) grows at the rate $\sqrt{n}$. By using variance estimates that are valid under global misspecification in place of the variance estimate $\hat{k}'_{\lambda^*_M}\hat{\Sigma}\hat{k}_{\lambda^*_M}$ in the CI construction, one can ensure that the resulting CI also remains valid under global misspecification. See Appendix D.2 for details.

REMARK 5.1. This framework can also be used to incorporate a priori restrictions on the magnitude of coefficients on control variables in an instrumental variables regression. Suppose that we have a set of controls w_i, that appear in the structural equation (21), so that $y_i = x_i'\theta + w_i'\gamma/\sqrt{n} + \epsilon_i$, and ϵ_i is uncorrelated with w_i as well as vector of instruments $\tilde{z}_i$. If one is willing to restrict the magnitude of the coefficient vector γ, so that $\|\gamma\| \le M$, then one can add w_i to the original vector of instruments $\tilde{z}_i$, $z_i = (\tilde{z}_i', w_i')'$. For example, if one is concerned with functional form misspecification, one can define the control variables to be higher order series terms. We then obtain the misspecified IV model with the set $\mathcal{C}$ given by (12), with $B = E[z_i w_i']$. Thus, we can interpret this model as a locally misspecified version of a model with w_i used as an excluded instrument.

REMARK 5.2. Instead of bounding the coefficient vector γ, one can alternatively bound the magnitude of the direct effect $z_{Ii}'\gamma$. If all instruments are potentially invalid, $z_{Ii} = z_i$, and one sets $\mathcal{C} = \{\gamma\colon E[(z_i'\gamma)^2] \le M\}$, then under homoscedasticity, this corresponds to the case discussed in Section 4.3.4, where the uncertainty from potential misspecification is exactly proportional to the asymptotic sampling uncertainty in $\hat{g}(\theta)$. Consequently, in this case the optimal sensitivity is the same as that given by the 2SLS estimator.

5.2 *Omitted variables bias in linear regression*

Specializing to the case where $z_i = x_i$, the misspecified IV model of Section 5.1 gives a misspecified linear regression model as a special case. This can be used to assess sensitivity of regression results to issues such as omitted variables bias. In particular, consider the linear regression model

$$y_i = x_i'\theta + w_i^* + \tilde{\varepsilon}_i, \quad Ex_i\tilde{\varepsilon}_i = 0,$$

where x_i and y_i are observed and w_i^* is a (possibly unobserved) omitted variable. Correlation between w_i^* and x_i will lead to omitted variables bias in the OLS regression of y_i on x_i. If w_i^* is unobserved, then we obtain our framework by making the assumption $\sqrt{n}Ew_i^*x_i \in \mathcal{C}$, for some set $\mathcal{C}$, and letting $\hat{g}(\theta) = \frac{1}{n}\sum_{i=1}^n x_i(y_i - x_i'\theta)$. This setup can also cover choosing between different sets of control variables. Suppose that $w_i^* = w_i'\gamma$, where w_i is a vector of observed control variables that the researcher is considering not including in the regression. If γ is unrestricted, then by the results in Section 4.3.2, the long regression of y_i on both x_i and w_i yields nearly optimal CIs. If one is willing to restrict the magnitude of γ, it is possible to tighten these CIs, with the setting reducing to that in Remark 5.1, with $z_i = (x_i, w_i')'$. The same framework can be used to incorporate selection bias by defining w_i^* to be the inverse Mills ratio term in the formula for $E[y_i \mid x_i, i \text{ observed}]$ in Heckman (1979).

5.3 *Functional form misspecification*

Our setup allows for misspecification in moment conditions arising from functional form misspecification. To apply our setup, one must relate this misspecification to the bounds $\mathcal{C}$ on the moment conditions at the true parameter value. One approach to bounding functional form misspecification is to use smoothness conditions from the nonparametric statistics literature, such as bounds on derivatives (see, e.g., Tsybakov (2009), for an introduction to this literature). Since these sets are typically convex (taking a convex combination of two functions that satisfy a given bound on a given derivative gives a function that also satisfies this bound), they typically lead to convex sets $\mathcal{C}$, so that our framework can be applied.

As a simple example, consider a nonparametric IV model with discrete covariates:

$$E\big[y_i - m(x_i) \mid z_i\big] = 0.$$

Suppose x takes values in the finite set $\mathcal{X} = \{\tilde{x}_1, \ldots, \tilde{x}_{N_x}\}$ and z_i takes values in the finite set $\mathcal{Z} = \{\tilde{z}_1, \ldots, \tilde{z}_{N_z}\}$. This setting was considered by Freyberger and Horowitz (2015), who place only nonparametric smoothness or shape restrictions on the unknown function m. To see the connection with our setting, we note that such restrictions can be interpreted as bounds on specification error from a parametric model. If one models these restrictions as local to a parametric family, one obtains our setting. In particular, let $m(x_i) = f(x_i, \theta_0) + n^{-1/2} r(x_i)$, $r \in \mathcal{R}$, where $\mathcal{R}$ is a nonparametric smoothness class. For example, if x_i is univariate, we can let $f(x_i, \theta) = \theta_1 + \theta_2 x_i$ and define $\mathcal{R}$ to be the class of functions with $r(0) = r'(0) = r''(0)$ and second derivative bounded by some constant M. This is equivalent to placing the bound $n^{-1/2}M$ on the second derivative of $m(\cdot)$, which corresponds to a Hölder smoothness class. We can then map this to a misspecified GMM model, with the jth element of the moment function given by $g_j(x_i, y_i, \theta) = (y_i - f(x_i, \theta_0))I(z_i = \tilde{z}_j)$ and jth element of the misspecification vector c given by $Er(x_i)I(z_i = \tilde{z}_j) = \sum_{\tilde{x} \in \mathcal{X}} r(\tilde{x})P(x_i = \tilde{x}, z_i = \tilde{z}_j)$. Stacking these equations, we see that $c = B\gamma$ where B is a matrix composed of the elements $P(x_i = \tilde{x}, z_i = \tilde{z}_j)$ and $\gamma = (r(\tilde{x}_1), \ldots, r(\tilde{x}_{N_x}))'$. As with the IV setting in Section 5.1, B is unknown, but can be replaced by a consistent estimate based on the sample analogue. So long as the set $\mathcal{R}$ is convex, we obtain convex restrictions on γ and therefore c, so that our framework applies.

This example brings up an important point about the interpretation of $h(\theta)$. If the object of interest is a functional of $m(x) = f(x, \theta_0) + n^{-1/2} r(x)$, then we will need to allow the object of interest $h(\cdot)$ to depend on the misspecification vector directly, as well as on θ. As discussed at the beginning of Section 2, this falls into a mild extension of our framework. Alternatively, under a suitable parametrization of f and r, it is often possible to define the object of interest to be function of θ alone. For example, if we are interested in the derivative $m'(x_0)$ at a particular point x_0 under a bound on the second derivative of $m(\cdot)$, we can let $f(x, \theta) = \theta_1 + \theta_2 x$ and define $\mathcal{R}$ to be the class of functions with $r(x_0) = r'(x_0) = r''(x_0) = 0$ and second derivative bounded by M. Then $m'(x_0) = \theta_2$.

5.4 *Treatment effect extrapolation*

Often, the average effect of a counterfactual policy on a particular subset of a population is of interest, and we would like to weaken the assumptions under which this effect is point-identified. We have available estimates $\hat{\tau} = (\hat{\tau}_1, \ldots, \hat{\tau}_m)'$ of the parameter τ, with $\hat{g}(\theta) = \hat{\tau} - A\theta$, and $g(\theta) = \tau - A\theta$ for a known matrix A. We would like to extrapolate from these estimates to learn about the parameter of interest $h(\theta_0) = H\theta_0$. The potential extrapolation bias is captured by the assumption that $g(\theta_0) \in \widetilde{\mathcal{C}}_n$, some convex set.

Note that, because the moment condition is linear in the parameter of interest, asymptotic validity of our CIs does not require that the set $\widetilde{\mathcal{C}}_n$ takes the form $\widetilde{\mathcal{C}}_n = \mathcal{C}/\sqrt{n}$. CIs given in equation (7) based on linear estimators of the form $\hat{h} = k'\hat{\tau}$ (such as minimum distance estimators), with $\mathcal{C} = \sqrt{n}\widetilde{\mathcal{C}}_n$ are valid under both local and global misspecification (i.e., under the assumption that the set $\widetilde{\mathcal{C}}_n$ is fixed as $n \to \infty$).

One example that falls into this setup are differences-in-differences designs when the parallel trends assumption is violated. Here, there are m time periods, with treatment taking place in period T_0. The $(m - T_0)$-vector θ corresponds to a vector of dynamic treatment effects on the treated, $A\theta = (0', \theta')'$, and $g(\theta_0)$ is a vector of trend differences between the treated and untreated, with $g(\theta_0) = 0$ if the parallel trends assumption holds. Rambachan and Roth (2019) built on the framework in this paper to develop CIs in this setting.

Another example that has been of recent interest involves nonseparable models with endogeneity. Under conditions in Imbens and Angrist (1994) and Heckman and Vytlacil (2005), instrumental variables estimates $\hat{\tau}_m$ with different instruments are consistent for average treatment effects for different subpopulations. A recent literature (Kowalski (2016), Brinch, Mogstad, and Wiswall (2017), Mogstad, Santos, and Torgovitsky (2018)) has focused on using assumptions on treatment effect heterogeneity to extrapolate these estimates to other populations. Our framework applies if these assumptions amount to placing the differences between the estimated treatment effects and the effect of interest in a known convex set.

6. Empirical application

This section illustrates the confidence intervals developed in Section 2 in an empirical application to automobile demand based on the data and model in Berry, Levinsohn, and Pakes (1995, BLP hereafter). We use the version of the model as implemented by Andrews, Gentzkow, and Shapiro (2017), who calculate the asymptotic bias of the GMM estimator with weighting matrix Σ^{-1} under local misspecification in this setting.[8]

6.1 *Model description and implementation*

In this model, the utility of consumer i from purchasing a vehicle j, relative to the outside option, is given by a random-coefficient logit model $U_{ij} = \sum_{k=1}^{K} x_{jk}(\beta_k + \sigma_k v_{ik}) -$

[8]The dataset for this empirical application has been downloaded from the Andrews, Gentzkow, and Shapiro (2017) replication files, available at https://doi.org/10.7910/DVN/LLARSN.

$\alpha p_j/y_i + \xi_j + \epsilon_{ij}$, where p_j is the price of the vehicle, x_{jk} the kth observed product characteristic, ξ_j is an unobserved product characteristic, and ϵ_{ij} is has an i.i.d. extreme value distribution. The income of consumer i is assumed to be log-normally distributed, $y_i = e^{m+\varsigma v_{i0}}$, where the mean m and the variance ς of log-income are assumed to be known and set to equal to estimates from the Current Population Survey. The unobservables $v_i = (v_{i0}, \ldots, v_{iK})$ are i.i.d. standard normal, while the distribution of the unobserved product characteristic ξ_j is unrestricted.

The marginal cost mc_j for producing vehicle j is given by $\log(mc_j) = w_j'\nu + \omega_j$, where w_j are observable characteristics, and ω_j is an unobservable characteristic. The full vector of model parameters is given by $\theta = (\sigma', \alpha, \beta', \nu')'$. Given this vector, and given a vector of unobservable characteristics, one can compute the market shares implied by utility maximization, which can be inverted to yield the unobservable characteristic as a function of θ, $\xi_j(\theta)$. One can similarly invert the unobserved cost component, writing it as a function of θ, $\omega_j(\theta)$, under the assumption that firms set prices to maximize profits in a Bertrand–Nash equilibrium. Given a vector z_{dj} of demand-side instruments, and a vector z_{sj} of supply-side instruments, this yields the moment condition $g(\theta) = E[\hat{\gamma}(\theta)]$, where

$$\hat{g}(\theta) = \frac{1}{n}\sum_{j=1}^{n}\begin{pmatrix} z_{dj}\xi_j(\theta) \\ z_{sj}\omega_j(\theta)\end{pmatrix}.$$

The BLP data spans the period 1971 to 1990, and includes information on essentially all $n = 999$ models sold during that period (for simplicity, we have suppressed the time dimension in the description above). There are 5 observable characteristics x_j: a constant, horsepower per 10 pounds of weight (HPWt), a dummy for whether air-conditioning is standard (Air), mileage per 10 dollars (MP$) defined as MPG over average gas price in a given year, and car size (Size), defined as length times width. The vector z_{dj} consists of x_j, plus the sum of x_j across models other than j produced by the same firm, and for rival firms. There are 6 cost variables w_j: a constant, log of HPWt, Air, log of MPG, log of Size, and a time trend. The vector z_{sj} consists of these variables, MP$, and the sums of w_j for own-firm products other than j, and for rival firms. After excluding collinear instruments, this gives a total of $d_g = 31$ instruments, 25 of which are excluded to identify $d_\theta = 17$ model parameters. The parameter of interest is average markup, $h(\theta) = \frac{1}{n}\sum_j (p_j - mc_j(\theta))/p_j$.

One may worry that some of these instruments are invalid, because elements of z_{dj} or z_{sj} may appear directly in the utility or cost function with the coefficient on the ℓth element given by $\delta_{d\ell}\gamma_{d\ell}/\sqrt{n}$ or $\delta_{s\ell}\gamma_{s\ell}/\sqrt{n}$, respectively. Here, $\delta_{d\ell}$ and $\delta_{s\ell}$ are scaling constants so that, given the sample size $n = 999$ at hand, $\gamma_{d\ell}$ has the interpretation that the consumer willingness to pay for one standard deviation change in the ℓth demand-side instrument $z_{dj\ell}$ is $\gamma_{d\ell}$% of the average 1980 car price, and changing the ℓth supply-side instrument $z_{sj\ell}$ by one standard deviation changes the marginal cost by $\gamma_{s\ell}$% of the average car price. Andrews, Gentzkow, and Shapiro (2017) used this scaling in their sensitivity analysis, and they discuss economic motivation for concerns about this form of misspecification. By way of comparison, the estimates of the parameters β and ν in the utility and cost function imply that consumers are on average willing to pay between 2.2

and 10.0% of the average car price for a standard deviation change in one of the included car characteristics, and that a standard deviation change in the included cost characteristics changes the marginal cost by between 3.8 and 11.1% of the average car price. We therefore interpret specifications of the set $\mathcal{C}$ that allow for $|\gamma_{s\ell}| \approx 1$–$2$ (or $|\gamma_{d\ell}| \approx 1$–$2$) as allowing for moderate amounts of misspecification in the ℓth supply-side (or demand-side) instrument.

We follow the implementation in Section 3. Given a set I of potentially invalid instruments, we follow Remark 3.1 and consider sets $\mathcal{C}$ of the form (12), with $\|\cdot\|$ corresponding to an ℓ_p norm with $p \in \{2, \infty\}$, and $B = \tilde{B}_I \cdot \#I^{1/p}$, where $\tilde{B}_I$ is given by the columns of

$$\tilde{B} = \begin{pmatrix} E\big[z_{dj} z_{dj}'\big] \operatorname{diag}(\delta_{d\ell}) & 0 \\ 0 & E\big[z_{sj} z_{sj}'\big] \operatorname{diag}(\delta_{d\ell}) \end{pmatrix},$$

and $\#I$ is the number of potentially invalid instruments. The scaling by $(\#I)^{1/p}$ ensures that the vector $\gamma = M(1, \ldots, 1)'$ is always included in the set. Andrews, Gentzkow, and Shapiro (2017) reported the sensitivity of the usual GMM estimator under this form of misspecification, considering misspecification in each instrument individually (so that I contains a single element), and setting $M = 1$. However, if one is concerned about the validity of several instruments, it is natural to allow I to contain all instruments the validity of which is questionable. In our analysis, we vary the set of potentially misspecified instruments. We also vary M in order to assess the sensitivity of conclusions to different amounts of misspecification. As we will see below, different choices of $\mathcal{C}$ lead to different sensitivities for the optimal estimator, and using the optimal sensitivity can reduce the width of the CI substantially relative to CIs based on the usual GMM estimator.

We use the estimate $\hat{\theta}_{\text{initial}}$ that corresponds to the GMM estimator based on the weight matrix that's optimal under correct specification, as reported in Andrews, Gentzkow, and Shapiro (2017), and the estimates $\hat{\Gamma}$, $\hat{H}$, and $\hat{\Sigma}$ are computed following Step 1 of the implementation.

6.2 *Results*

To illustrate that using the sensitivity that is optimal under local misspecification can yield substantially tighter CIs, Figure 1 plots the confidence intervals based on the optimal sensitivity, as well as those based on $\hat{\theta}_{\text{initial}}$ under different sets I of potentially invalid instruments and ℓ_2 constraints on γ. It is clear from the figure that using the optimal sensitivity yields substantially tighter confidence intervals, relative to simply adjusting the usual CI by using the critical value $\text{cv}_\alpha(\cdot)$ to take into account the potential bias of $h(\hat{\theta}_{\text{initial}})$, by as much as a factor of 3.4. The intuitive reason for this is that by adjusting the sensitivity of the estimator, it is possible to substantially reduce its bias at little cost in terms of an increase in variance. Thus, for example, while the CI for the average markup based on the estimate $\hat{\theta}_{\text{initial}}$ is essentially too wide to be informative when the set of potentially invalid instruments corresponds to all excluded instruments, the CI based on the optimal sensitivity, $[46.0, 66.0]\%$, is still quite tight.

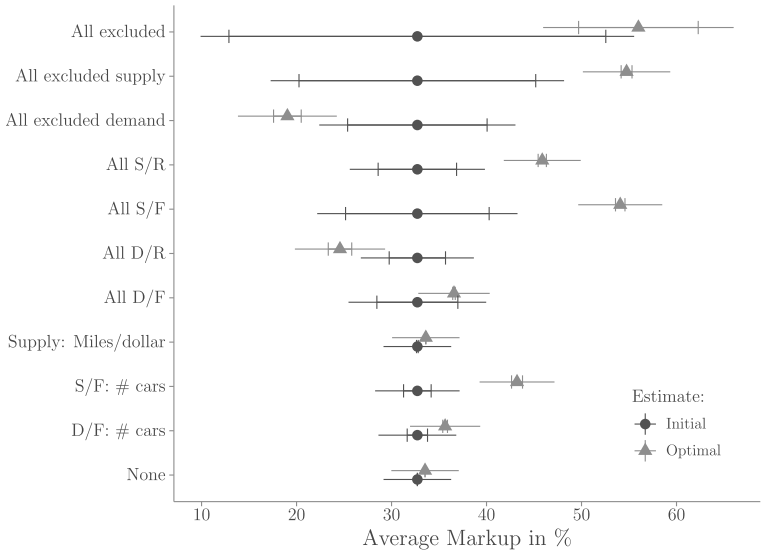

FIGURE 1. Confidence intervals under ℓ_2 misspecification and $M = 1$ in the application to Berry, Levinsohn, and Pakes (1995).

If a researcher is ex ante unsure what form of misspecification one should worry about, as a sensitivity check, it is useful to consider the effects of different forms of misspecification. In Figure 2, we plot the optimal confidence intervals for different subsets of invalid instruments, under both ℓ_2 and ℓ_∞ norms for γ. Although the choice of norm matters when the number of potentially misspecified instruments is greater than one,

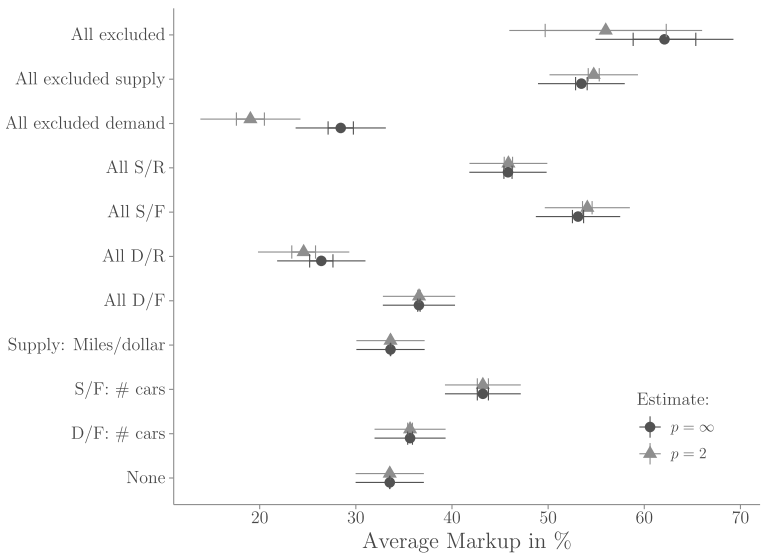

FIGURE 2. Optimal Confidence intervals under ℓ_2, and ℓ_∞ misspecification and $M = 1$ in the application to Berry, Levinsohn, and Pakes (1995).

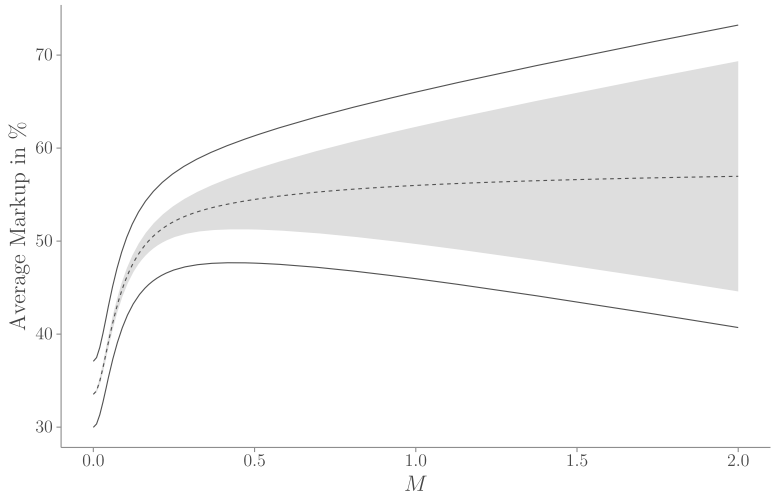

FIGURE 3. Optimal confidence intervals under ℓ_2 misspecification the application to Berry, Levinsohn, and Pakes (1995) as a function of misspecification parameter M, when all excluded instruments are allowed to be potentially invalid.

the results are qualitatively similar. Comparing the results for different choices of the set of potentially invalid instruments suggests that allowing supply-side instruments to be invalid generally increases the average markup estimate, while allowing demand-side instruments to be invalid has the opposite effect.

As it may be ex ante unclear what magnitude of misspecification is reasonable to allow for, as discussed in Section 3, it is useful to plot the optimal CI for multiple choices of M. We do this in Figure 3 for $p = 2$, and we allow all excluded instruments to be potentially invalid. One can see that while the CI is unstable for values of M smaller than about 0.4, for larger values of M, the estimate is quite stable and equal to about 50%. Even at $M = 2$, one rejects the hypothesis that the optimal markup is equal to the initial estimate $h(\hat{\theta}_{\text{initial}}) = 32.7\%$. This suggests that ignoring misspecification in the BLP model likely leads to a downward bias in the estimate of the average markup. At the same time, it is possible to obtain reasonably tight CIs for the average markup even under a moderate amount of misspecification.

The J-statistic for testing the hypothesis that all moments are correctly specified equals 426.7. Consequently, the hypothesis is rejected at the usual significance levels. Furthermore, it can be seen from Figure 2 that the CIs for "all excluded" (that allow all excluded instruments to be invalid at $M = 1$), and "all excluded demand" (that assume validity of supply-side instruments) do not overlap. This implies that either the misspecification in the demand-side instruments must be greater than 1% of the average care price ($M = 1$), or else the supply-side instruments must also be invalid. Table 1 implements the specification test from Appendix B that gives lower CI $[M_{\text{min}}, \infty]$ for M. The results suggest that if one assumes only a subset of the instruments is invalid, the misspecification in the potentially invalid instruments must be quite large. For example, if we assume that all instruments are valid except potentially the demand-side instruments based on rival firms' product characteristics, then the misspecification in these

TABLE 1. J-Test of overidentifying restrictions in the application to Berry, Levinsohn, and Pakes (1995) under different forms of ℓ_p misspecification.

Instrument set	$p=2$	$p=\infty$
D/F: # cars	10.21	10.21
S/F: # cars	15.00	15.00
Supply: Miles/dollar	16.31	16.31
All D/F	2.71	2.71
All D/R	5.36	5.55
All S/F	2.54	2.56
All S/R	4.06	6.84
All excluded demand	1.80	1.97
All excluded supply	1.60	1.72
All excluded	1.13	2.56

Note: The table gives the minimum value of M such that the test of overidentifying restrictions has p-value equal to 0.05. "D/F": Demand-side instrument based on characteristics of other cars produced by the same firm. "S/F": Supply-side instrument based on characteristics of other cars produced by the same firm. "D/R": Demand-side instrument based on characteristics of cars produced by rivals. "S/R": Supply-side instrument based on characteristics of cars produced by rivals. "All excluded": All excluded instruments are potentially invalid.

instruments must be greater than $M = 5.36$. If we allow all instruments to be invalid, then $M \geq 1.13$.

Finally, to illustrate the implication of Theorem 4.1 that one cannot substantively improve upon the CIs that we construct, we calculate the efficiency bound κ_* for these CIs in Table 2. The table shows that the bound is at least as high as the efficiency bound for the usual CI under correct specification (given in (19) and equal to 84.99% at $\alpha = 0.05$). Thus, the asymptotic scope for improvement over the CIs reported in Figure 2 at

TABLE 2. Efficiency bounds (in %) for one and two-sided 95% confidence intervals at $c = 0$ under ℓ_p misspecification in the application to Berry, Levinsohn, and Pakes (1995).

	Two-sided		One-sided	
Instrument set	$p=2$	$p=\infty$	$p=2$	$p=\infty$
D/F: # cars	85.9	85.9	100.0	100.0
S/F: # cars	90.1	90.1	99.8	99.8
Supply: Miles/dollar	85.0	85.0	100.0	100.0
All D/F	85.5	85.7	100.0	100.0
All D/R	94.8	95.3	93.9	95.3
All S/F	88.6	89.1	99.7	99.7
All S/R	89.4	89.2	98.5	99.5
All excluded demand	95.4	96.4	95.0	97.3
All excluded supply	90.3	90.1	98.2	99.6
All excluded	97.0	97.5	99.5	98.2

Note: For two-sided confidence intervals, the table calculates the ratio of the expected length of a 95% confidence interval that minimizes its length at $c = 0$ relative to the length of the CI in (16), given in (18). For one-sided confidence intervals, the table calculates an analogous bound, given in Appendix C.6, when the confidence interval optimizes the 80% quantile of excess length. Instrument set labels are describe in notes to Table 1.

particular values of θ and $c = 0$ is even smaller than the scope for improvement over the usual CI at particular values of θ under correct specification.

References

Andrews, D. W. K. and P. Guggenberger (2009), "Hybrid and size-corrected subsampling methods." *Econometrica*, 77 (3), 721–762. [79, 93]

Andrews, I., M. Gentzkow, and J. M. Shapiro (2017), "Measuring the sensitivity of parameter estimates to sample statistics." *Quarterly Journal of Economics*, 132 (4), 1553–1592. [78, 79, 80, 82, 85, 96, 99, 100, 101]

Andrews, I., M. Gentzkow, and J. M. Shapiro (2020), "On the informativeness of descriptive statistics for structural estimates." *Econometrica* (forthcoming). [80, 94]

Armstrong, T. B. and M. Kolesár (2018), "Optimal inference in a class of regression models." *Econometrica*, 86 (2), 655–683. [79, 90, 95]

Armstrong, T. B. and M. Kolesár (2021), "Supplement to 'Sensitivity analysis using approximate moment condition models'." *Quantitative Economics Supplemental Material*, 12, https://doi.org/10.3982/QE1609. [81]

Berkowitz, D., M. Caner, and Y. Fang (2012), "The validity of instruments revisited." *Journal of Econometrics*, 166 (2), 255–266. [80]

Berry, S. T., J. Levinsohn, and A. Pakes (1995), "Automobile prices in market equilibrium." *Econometrica*, 63 (4), 841–890. [79, 99, 102, 103, 104]

Bonhomme, S. and M. Weidner (2020), "Minimizing sensitivity to model misspecification." arXiv:1807.02161. [80, 86]

Brinch, C. N., M. Mogstad, and M. Wiswall (2017), "Beyond LATE with a discrete instrument." *Journal of Political Economy*, 125 (4), 985–1039. [99]

Bugni, F. A. and T. Ura (2019), "Inference in dynamic discrete choice problems under local misspecification." *Quantitative Economics*, 10 (1), 67–103. [80]

Cai, T. T. and M. G. Low (2004), "An adaptation theory for nonparametric confidence intervals." *The Annals of Statistics*, 32 (5), 1805–1840. [79]

Chamberlain, G. (1987), "Asymptotic efficiency in estimation with conditional moment restrictions." *Journal of Econometrics*, 34 (3), 305–334. [80]

Conley, T. G., C. B. Hansen, and P. E. Rossi (2012), "Plausibly exogenous." *The Review of Economics and Statistics*, 94 (1), 260–272. [80, 85, 87, 96]

Cressie, N. and T. R. C. Read (1984), "Multinomial goodness-of-fit tests." *Journal of the Royal Statistical Society. Series B (Methodological)*, 46 (3), 440–464. [94]

DiTraglia, F. J. (2016), "Using invalid instruments on purpose: Focused moment selection and averaging for GMM." *Journal of Econometrics*, 195 (2), 187–208. [79, 93]

Donoho, D. L. (1994), "Statistical estimation and optimal recovery." *The Annals of Statistics*, 22 (1), 238–270. [79, 83, 89]

Duflo, E., M. Greenstone, R. Pande, and N. Ryan (2018), "The value of regulatory discretion: Estimates from environmental inspections in India." *Econometrica*, 86 (6), 2123–2160. [78]

Efron, B., T. Hastie, I. M. Johnstone, and R. J. Tibshirani (2004), "Least angle regression." *The Annals of Statistics*, 32 (2), 407–451. [79, 86]

Freyberger, J. and J. L. Horowitz (2015), "Identification and shape restrictions in nonparametric instrumental variables estimation." *Journal of Econometrics*, 189 (1), 41–53. [98]

Gayle, G.-L. and A. Shephard (2019), "Optimal taxation, marriage, home production, and family labor supply." *Econometrica*, 87 (1), 291–326. [78]

Guggenberger, P. (2012), "On the asymptotic size distortion of tests when instruments locally violate the exogeneity assumption." *Econometric Theory*, 28 (2), 387–421. [80]

Hahn, J. and J. A. Hausman (2005), "IV estimation with valid and invalid instruments." *Annales d'Économie et de Statistique*, 79/80, 25–57. [96]

Hall, A. R. and A. Inoue (2003), "The large sample behaviour of the generalized method of moments estimator in misspecified models." *Journal of Econometrics*, 114 (2), 361–394. [87]

Hansen, L. P. (1985), "A method for calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators." *Journal of Econometrics*, 30 (1–2), 203–238. [80]

Hansen, L. P. and T. J. Sargent (2008), *Robustness*. Princeton University Press, Princeton, NJ. [86]

Heckman, J. J. (1979), "Sample selection bias as a specification error." *Econometrica*, 47 (1), 153–161. [97]

Heckman, J. J. and E. Vytlacil (2005), "Structural equations, treatment effects and econometric policy evaluation." *Econometrica*, 73 (3), 669–738. [99]

Hong, H., A. Mahajan, and D. Nekipelov (2015), "Extremum estimation and numerical derivatives." *Journal of Econometrics*, 188 (1), 250–263. [85]

Huber, P. J. and E. M. Ronchetti (2009), *Robust Statistics*, second edition. John Wiley & Sons, Hoboken, NJ. [80]

Imbens, G. W. and J. D. Angrist (1994), "Identification and estimation of local average treatment effects." *Econometrica*, 62 (2), 467–475. [99]

Imbens, G. W. and C. F. Manski (2004), "Confidence intervals for partially identified parameters." *Econometrica*, 72 (6), 1845–1857. [81, 91]

Joshi, V. M. (1969), "Admissibility of the usual confidence sets for the mean of a univariate or bivariate normal population." *The Annals of Mathematical Statistics*, 40 (3), 1042–1067. [90]

Kitamura, Y., T. Otsu, and K. Evdokimov (2013), "Robustness, infinitesimal neighborhoods, and moment restrictions." *Econometrica*, 81 (3), 1185–1201. [80, 94]

Kowalski, A. E. (2016), "Doing more when you're running LATE: Applying marginal treatment effect methods to examine treatment effect heterogeneity in experiments." Working Paper 22363, National Bureau of Economic Research. [99]

Low, M. G. (1995), "Bias-variance tradeoffs in functional estimation problems." *The Annals of Statistics*, 23 (3), 824–835. [89]

Masten, M. A. and A. Poirier (2020), "Salvaging falsified instrumental variable models." arXiv:1812.11598. [86]

McCloskey, A. (2020), "Asymptotically uniform tests after consistent model selection in the linear regression model." *Journal of Business & Economic Statistics*, 38 (4) 810–825. [79, 93]

Mogstad, M., A. Santos, and A. Torgovitsky (2018), "Using instrumental variables for inference about policy relevant treatment effects." *Econometrica*, 86 (5), 1589–1619. [99]

Moon, H. R. and F. Schorfheide (2009), "Estimation with overidentifying inequality moment conditions." *Journal of Econometrics*, 153 (2), 136–154. [95]

Mukhin, Y. (2018), "Sensitivity of regular estimators." arXiv:1805.08883. [80]

Newey, W. K. (1985), "Generalized method of moments specification testing." *Journal of Econometrics*, 29 (3), 229–256. [80, 82]

Newey, W. K. (1990), "Semiparametric efficiency bounds." *Journal of Applied Econometrics*, 5 (2), 99–135. [82]

Newey, W. K. and D. L. McFadden (1994), "Large sample estimation and hypothesis testing." In *Handbook of Econometrics*, Vol. 4 (D. L. McFadden and R. F. Engle, eds.), Chapter 36, 2111–2245, Elsevier. [80, 82, 85]

Pratt, J. W. (1961), "Length of confidence intervals." *Journal of the American Statistical Association*, 56 (295), 549–567. [90]

Rambachan, A. and J. Roth (2019), "An honest approach to parallel trends." Unpublished manuscript, Harvard University. [99]

Rosset, S. and J. Zhu (2007), "Piecewise linear regularized solution paths." *The Annals of Statistics*, 35 (3), 1012–1030. [79]

Sacks, J. and D. Ylvisaker (1978), "Linear estimation for approximately linear models." *The Annals of Statistics*, 6 (5), 1122–1137. [79, 88]

Tsybakov, A. B. (2009), *Introduction to Nonparametric Estimation*. Springer, New York, NY. [98]

van der Vaart, A. W. (1998), *Asymptotic Statistics*. Cambridge University Press, New York, NY. [80, 92, 93]

Co-editor Andres Santos handled this manuscript.

Manuscript received 21 April, 2020; final version accepted 1 September, 2020; available online 16 September, 2020.